involve

a journal of mathematics

msp

2020 vol. 13, no. 4

involve

msp.org/involve

INVOLVE YOUR STUDENTS IN RESEARCH

Involve showcases and encourages high-quality mathematical research involving students from all academic levels. The editorial board consists of mathematical scientists committed to nurturing student participation in research. Bridging the gap between the extremes of purely undergraduate research journals and mainstream research journals, *Involve* provides a venue to mathematicians wishing to encourage the creative involvement of students.

Cover: Alex Scorpan

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2020 is US $205/year for the electronic version, and $275/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

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

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

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

INVOLVE 13:4 (2020)
https://doi.org/10.2140/involve.2020.13.541

Structure constants of $\mathcal{U}(\mathfrak{sl}_2)$

Alexia Gourley and Christopher Kennedy

(Communicated by Joseph A. Gallian)

We derive a recursive formula for the structure constants for the universal enveloping algebra of $\mathfrak{sl}_2(\mathbb{K})$, where $\mathbb{K}$ is an algebraically closed field of characteristic zero.

1. Introduction

Given an algebra $\mathcal{A}$ over a field $\mathbb{K}$ with basis $\{v_k \mid k \in I\}$ indexed by some set I, the product of any two basis vectors can be expressed in terms of the basis; i.e., we can write $v_i v_j$ as $\sum_{k\in I} \gamma_{ijk} v_k$ for some scalars $\gamma_{ijk} \in \mathbb{K}$. The γ_{ijk} are referred to as structure constants and theoretically, we can discern properties of the algebra from identities satisfied, or not satisfied, by these scalars. For example, the algebra $\mathcal{A}$ will be commutative if and only if $\gamma_{ijk} = \gamma_{jik}$ for all i and j. From a practical standpoint, structure constants are of the most use for smaller algebras, i.e., finite-dimensional. For an infinite-dimensional algebra, computing the structure constants is a daunting task, especially since the structure constants with respect to one basis are likely to differ wildly from the structure constants with respect to another basis.

With that being said, the case we consider here is almost ideal for this type of problem. The Lie algebra $\mathfrak{sl}_2(\mathbb{K})$ is ubiquitous in Lie theory, small (three-dimensional) and easily described through its structure constants. While its universal enveloping algebra, $\mathcal{U}(\mathfrak{sl}_2(\mathbb{K}))$ is infinite-dimensional, thanks to the very powerful Poincaré-Birkhoff-Witt theorem, it has a canonical basis generated by the basis of the underlying $\mathfrak{sl}_2(\mathbb{K})$. This makes computing the structure constants somewhat tractable. While the mathematics contained herein is not tremendously complicated, and relies heavily on induction, the result itself is quite interesting.

2. Statement of the problem

For the purpose of self-containment, we give several initial definitions and results, the latter without proof. We refer the interested reader to either [Erdmann and Wildon 2006] for a general introduction to Lie algebras and enveloping algebras or [Carter 2005; Humphreys 1972] for a more in-depth development of the theory.

MSC2010: 17B05.
Keywords: structure constants, universal enveloping algebra, nonassociative algebra, Lie algebras.

Definition 1. A *Lie algebra* is a vector space $\mathfrak{L}$ over a field $\mathbb{K}$ equipped with a bilinear product, called the bracket, $[\,\cdot\,,\cdot\,]:\mathfrak{L}\times\mathfrak{L}\to\mathfrak{L}$ satisfying

- (anticommutativity) $[x,x]=0$ for all $x\in\mathfrak{L}$;
- (Jacobi identity) $[[x,[y,z]]+[y,[z,x]]+[z,[x,y]]=0$ for all $x,y,z\in\mathfrak{L}$.

If $\mathbb{K}$ has characteristic other than 2, by applying bilinearity to $[x+y,x+y]=0$ the immediate consequence is that $[x,y]=-[y,x]$; that is, we can reverse the order of the bracket at the cost of a negative.

Given an associative algebra $\mathcal{A}$, there is an associated Lie algebra $\mathcal{A}^-$ with the same underlying vector space equipped with the *commutator bracket*, $[x,y]:=xy-yx$ for all $x,y\in\mathcal{A}^-$. For example, the associative algebra $\mathrm{Mat}_2(\mathbb{K})$ of 2×2-matrices can be equipped with the commutator bracket to yield the *general linear algebra*, $\mathfrak{gl}_2(\mathbb{K})$. Here, the bracket of two matrices A and B is just $[A,B]=AB-BA$, with AB and BA being the regular matrix product. Because AB rarely equals BA for arbitrary matrices, the bracket is usually nonzero. A crucial, albeit small, example of a Lie algebra is $\mathfrak{sl}_2(\mathbb{K})$, which arises as a Lie subalgebra of $\mathfrak{gl}_2(\mathbb{K})$.

Definition 2. The *special linear Lie algebra* of 2×2 matrices, $\mathfrak{sl}_2(\mathbb{K})$, is the Lie algebra consisting of all 2×2 matrices with trace zero equipped with the commutator bracket.

The field $\mathbb{K}$ does not play an important role in what follows, and consequently will be omitted; e.g., $\mathfrak{sl}_2$ instead of $\mathfrak{sl}_2(\mathbb{K})$. The Lie algebra $\mathfrak{sl}_2$ is three-dimensional with a natural basis consisting of the matrices

$$e=\begin{bmatrix}0&1\\0&0\end{bmatrix},\quad f=\begin{bmatrix}0&0\\1&0\end{bmatrix},\quad h=\begin{bmatrix}1&0\\0&-1\end{bmatrix}.$$

The bracket of any of these with itself is zero by anticommutativity and the mixed brackets give the relations

$$[e,f]=h,\quad [h,e]=2e,\quad [h,f]=-2f. \tag{1}$$

A natural question to ask is whether a Lie algebra $\mathfrak{L}$ arises as the associated Lie algebra for some associative algebra, $\mathcal{A}$. That is, given $\mathfrak{L}$, does there exist an $\mathcal{A}$ such that $\mathfrak{L}=\mathcal{A}^-$? In certain cases the answer is yes (e.g., if $\mathfrak{L}$ is finite-dimensional), but in general the answer is no. That being said, we can often find an associative algebra $\mathcal{A}$ such that $\mathcal{L}$ is a subalgebra of $\mathcal{A}^-$; that is, $\mathfrak{L}\subseteq\mathcal{A}^-$. Such an algebra $\mathcal{A}$ is said to be an *enveloping algebra* of $\mathfrak{L}$. A Lie algebra can possess many distinct enveloping algebras, but the following theorem asserts the existence of a powerful and special one.

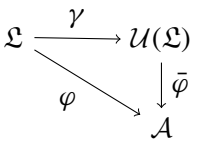
Theorem 3 (Poincaré–Birkhoff–Witt). *Let $\mathfrak{L}$ be a Lie algebra. Then there exists an associative algebra $\mathcal{U}(\mathfrak{L})$, called the* ***universal enveloping algebra****, satisfying*:

(1) *$\mathfrak{L}$ can be identified with a subalgebra of $\mathcal{U}(\mathfrak{L})$, where the bracket on $\mathfrak{L}$ is given by $[x,y]=xy-yx$.*

(2) *If $\varphi : \mathfrak{L} \to \mathcal{A}^-$ is any Lie algebra homomorphism, there is a unique associative algebra homomorphism*

$$\bar{\varphi} : \mathcal{U}(\mathfrak{L}) \to \mathcal{A}$$

such that the following diagram commutes:



Moreover, if $\mathfrak{L}$ has a basis $\{v_i \mid i \in I\}$, where I is equipped with a total order $<$, then $\mathcal{U}(\mathfrak{L})$ has a basis of all monomials of the form $v_{i_1}^{n_1} v_{i_2}^{n_2} \cdots v_{i_m}^{n_m}$, where $i_1 < i_2 < \cdots < i_m$ and $n_j \in \mathbb{N}$.

While a discussion of homomorphisms is beyond the scope of this paper, heuristically, what Theorem 3 is saying is that there is one enveloping algebra for $\mathfrak{L}$ that trumps all others — any enveloping algebra for a given $\mathfrak{L}$ can be derived from $\mathcal{U}(\mathfrak{L})$. In theory, if we know $\mathcal{U}(\mathfrak{L})$, we can determine any and all enveloping algebras of $\mathfrak{L}$.

In this paper, we are concerned with the universal enveloping algebra of $\mathfrak{sl}_2$. Since $\mathfrak{sl}_2$ is three-dimensional, we can order the canonical basis by $f < h < e$. Note that this ordering is arbitrary in the sense that there is nothing "larger" about e as compared to h and so forth. Really, it is just a reflection that f is lower triangular and we move "up" to h as a diagonal matrix and then farther "up" to e as upper triangular. With this convention set, Theorem 3 provides a basis of $\mathcal{U}(\mathfrak{sl}_2)$ consisting of all elements of the form $f^r h^s e^t$, where $r, s, t \in \mathbb{N}$.

For the sake of completeness, we give the formal definition of structure constants.

Definition 4. Let $\mathcal{A}$ be an algebra (not necessarily associative) with basis $\{v_k \mid k \in I\}$. The *structure constants* of $\mathcal{A}$ with respect to the basis $\{v_k \mid k \in I\}$ are scalars γ_{ijk}, where

$$v_i \cdot v_j = \sum_{k \in I} \gamma_{ijk} v_k,$$

where all but finitely many of the γ_{ijk} are zero.

With all this in place, we state the main problem. Note that while in Definition 4 the basis vectors are indexed by a single value, the PBW basis vectors $f^r h^s e^t$ are naturally indexed by the triple (r, s, t).

Main Problem. Given the Poincaré–Birkhoff–Witt basis $\{f^r h^s e^t \mid r, s, t \in \mathbb{N}\}$ of $\mathcal{U}(\mathfrak{sl}_2)$, what are the structure constants γ_{ijkl} such that

$$f^r h^s e^t \cdot f^u h^v e^w = \sum_{l \in L} \gamma_{ijkl} f^i h^j e^k \tag{2}$$

for some indexing set L?

Ideally, we would determine the constants explicitly. As we will see in Section 5, more realistically we develop a recursive formula for the γ_{ijkl} depending on the powers on the left-hand side of the equation.

3. Commutation identities: initial cases

To determine this, we need to use various commutation identities to reorder the components on the left-hand side of (2). In $\mathcal{U}(\mathfrak{sl}_2)$, the bracket identities from (1) give

$$[e, f] = ef - fe = h, \tag{3}$$

$$[h, e] = he - eh = 2e, \tag{4}$$

$$[h, f] = hf - fh = -2f. \tag{5}$$

These allow us to reorder the products of e, f, and h:

$$ef = fe + h, \tag{6}$$

$$eh = he - 2e, \tag{7}$$

$$hf = fh - 2f. \tag{8}$$

We make heavy use of mathematical induction in what follows and consequently begin by establishing several base cases.

Proposition 5. *Given a PBW basis element $f^r h^s e^t$, we have*

$$(f^r h^s e^t) f = f^{r+1}(h-2)^s e^t + t f^r h^{s+1} e^{t-1} - t(t-1) f^r h^s e^{t-1}. \tag{9}$$

Proof. To establish (9), we begin by determining expressions for $e^t f$ and $h^s f$ in terms of the PBW basis. Specifically, for the former we claim that

$$e^t f = f e^t + t h e^{t-1} - t(t-1)e^{t-1}. \tag{10}$$

If $t = 1$ in (10), it simplifies to (6) from earlier. Assuming the result for t, then

$$\begin{aligned}
(e^{t+1} f) &= e(e^t f) = e(f e^t + t h e^{t-1} - t(t-1)e^{t-1}) \\
&= (ef)e^t + t(eh)e^{t-1} - t(t-1)e^t \\
&= (fe + h)e^t + t(he - 2e)e^{t-1} - t(t-1)e^t \\
&= f e^{t+1} + h e^t + t h e^t - 2t e^t - t(t-1)e^t \\
&= f e^{t+1} + (t+1) h e^t - [2t + t(t-1)]e^t \\
&= f e^{t+1} + (t+1) h e^t - t(t+1)e^t.
\end{aligned}$$

Using a similar argument for $h^s f$, we claim

$$h^s f = f(h-2)^s. \tag{11}$$

If we take $s = 1$, it becomes (8) provided we distribute the f on the left. Assuming the result for s, we have

$$\begin{aligned}(h^{s+1} f) &= h(h^s f) = h(f(h-2)^s) = (hf)(h-2)^s \\ &= (fh-2f)(h-2)^s = f(h-2)(h-2)^s = f(h-2)^{s+1}.\end{aligned}$$

Applying (10) and (11) to $(f^r h^s e^t) f$, we have

$$\begin{aligned}(f^r h^s)(e^t f) &= (f^r h^s)(fe^t + the^{t-1} - t(t-1)e^{t-1}) \\ &= f^r(h^s f)e^t + tf^r h^{s+1} e^{t-1} - t(t-1) f^r h^s e^{t-1} \\ &= f^r(f(h-2)^s)e^t + tf^r h^{s+1} e^{t-1} - t(t-1) f^r h^s e^{t-1} \\ &= f^{r+1}(h-2)^s e^t + tf^r h^{s+1} e^{t-1} - t(t-1) f^r h^s e^{t-1}. \qquad \square\end{aligned}$$

It should be noted that (11) is not technically in PBW format, although it is close. While $\mathcal{U}(\mathfrak{sl}_2)$ is highly noncommutative, elements of the field $\mathbb{K}$ commute with everything. In particular, the scalar 2 commutes with h and the term $(h-2)^s$ can be expanded using the binomial theorem. Doing so yields the product in (11) in terms of the PBW basis:

$$(f^r h^s e^t) f = \left(\sum_{k=0}^{s} \binom{s}{k} (-2)^k f^{r+1} h^{s-k} e^t \right) + tf^r h^{s+1} e^{t-1} - t(t-1) f^r h^s e^{t-1}. \quad (12)$$

Obviously the equation in Proposition 5 is more compact.

Next we turn to pushing an h through a generic PBW basis vector.

Proposition 6. *Given a PBW basis element $f^r h^s e^t$, we have*

$$(f^r h^s e^t)h = f^r h^{s+1} e^t - 2tf^r h^s e^t. \quad (13)$$

Proof. As with the preceding proof, we need to establish the identity

$$e^t h = he^t - 2te^t = (h-2t)e^t. \quad (14)$$

If $t = 1$, this is (7). Assuming the result for t, we have

$$\begin{aligned}e^{t+1} h &= e(e^t h) = e(he^t - 2te^t) = (eh)e^t - 2te^{t+1} \\ &= (he-2e)e^t - 2te^{t+1} = he^{t+1} - 2(t+1)e^{t+1}.\end{aligned}$$

Applying this to (13), we have

$$\begin{aligned}(f^r h^s e^t)h &= f^r h^s (e^t h) = f^r h^s (h-2t)e^t \\ &= f^r h^{s+1} e^t - 2tf^r h^s e^t. \qquad \square\end{aligned}$$

4. Commutation identities: secondary cases

Now, we turn our attention to the products $e^t h^v$ and $h^s f^u$.

Proposition 7. *The quantities $e^t h^v$ and $h^s f^u$ can be expressed as*

$$e^t h^v = (h-2t)^v e^t = \sum_{k=0}^{v} \binom{v}{k} (-2t)^k h^{v-k} e^t, \tag{15}$$

$$h^s f^u = f^u (h-2u)^s = \sum_{k=0}^{u} \binom{u}{k} (-2u)^k f^u h^{u-k}. \tag{16}$$

Proof. The far right-hand sides of both equations result from applying the binomial theorem to centers of their respective equations, so we focus on establishing the left-hand equalities. Moreover, the proof of the second is symmetric to the first and is omitted. For the first, we proceed by induction on v with t held constant. If $v = 1$, we have $e^t h = (h-2t)e^t$, which was established in Proposition 6. Assuming the result for $e^t h^v$, we have

$$\begin{aligned} e^t h^{v+1} &= (e^t h^v)h = (h-2t)^v e^t h \\ &= (h-2t)^v (h-2t) e^t = (h-2t)^{v+1} e^t. \end{aligned}$$ □

Our next concern is how to express an element of the form $e^t f^u$ in terms of the PBW basis, with the case $u = 1$ simply being (10). This is decidedly more complicated than the cases in Proposition 7. If we consider $e^t h^v$, for example, we are in essence repeatedly using (7): $eh = he - 2e$. We can think of this as reordering eh as a sum of elements of the form $h^i e$, where the power of e remains constant and the powers of h decrease. This is precisely what happens with $e^t h^v$ and symmetrically with $h^s f^u$.

On the other hand, when we wish to reorder $e^t f^u$, we need to repeatedly use (6): $ef = fe + h$. Here, the h-term is essentially new. Successive applications will generate more powers of h, which will interact in turn with the various powers of f. Consequently, more and more terms are spawned as we push successive powers of f past successive powers of e. Consequently, it becomes logistically impossible to express $e^t f^u$ explicitly in the PBW basis, but it is quite natural to be done recursively.

Theorem 8. *The element $e^t f^u$ can be expressed in the form*

$$\sum_{i=0}^{u} \sum_{l=0}^{i} \gamma_{u,i,l} f^{u-i} h^{i-l} e^{t-i}$$

for scalars $\gamma_{u,i,l}$. Moreover for fixed natural numbers a and b with $b \le a$, we have the following recursive formula for the coefficient of $f^{u+1-a} h^{a-b} e^{t-a}$ in terms of the coefficients $\gamma_{u,i,l}$:

$$\begin{aligned} \gamma_{u+1,a,b} = \Bigg(\sum_{l+k=0}^{b} \binom{a-l}{k} (-2)^k \gamma_{u,a,l} + \gamma_{u,a-1,b}(t-a+1) \Bigg) \\ - \gamma_{u,a-1,b-1}(t-a)(t-a-1). \end{aligned} \tag{17}$$

Proof. We first establish the claim that we can express $e^t f^u$ as a linear combination of terms of the form $f^{u-i}h^{i-l}e^{t-i}$ for various i and l, and not surprisingly we proceed by induction on u. We have established $u=1$ in (10). The general approach is to use the induction hypothesis and compute $e^t f^{u+1} = (e^t f^u)f$ and then do considerable rearrangement and reindexing along with applying Propositions 5 and 6. In general, $\gamma_{u,i,l}$ is the coefficient of $f^{u-i}h^{i-l}e^{t-i}$ with $i \le u$ and $l \le i$. In what follows, we assign a value of 0 to any scalar $\gamma_{u,i,l}$ where $i > u$ or $l > i$ since no such terms exist. Additionally, if $i < 0$ or $l < 0$, we make the same assignment:

$$\begin{aligned}
(e^t f^u)f &= \sum_{i=0}^{u}\sum_{l=0}^{i}\gamma_{u,i,l}f^{u-i}h^{i-l}e^{t-i}f \\
&= \sum_{i=0}^{u}\sum_{l=0}^{i}\gamma_{u,i,l}f^{u-i}h^{i-l}(fe^{t-i}+(t-i)he^{t-i-1}-(t-i)(t-i-1)e^{t-i-1}) \\
&= \sum_{i=1}^{u}\sum_{l=0}^{i}\gamma_{u,i,l}f^{u-i}(h^{i-l}f)e^{t-i}+\sum_{i=0}^{u}\sum_{l=0}^{i}\gamma_{u,i,l}(t-i)f^{u-i}h^{i-l+1}e^{t-i-1} \\
&\quad -\sum_{i=0}^{u}\sum_{l=0}^{i}\gamma_{u,i,l}(t-i)(t-i-1)f^{u-i}h^{i-l}e^{t-i-1} \\
&= \sum_{i=0}^{u}\sum_{l=0}^{i}\gamma_{u,i,l}f^{u+1-i}(h-2)^{i-l}e^{t-i}+\sum_{i=0}^{u}\sum_{l=0}^{i}\gamma_{u,i,l}(t-i)f^{u-i}h^{i-l+1}e^{t-i-1} \\
&\quad -\sum_{i=0}^{u}\sum_{l=0}^{i}\gamma_{u,i,l}(t-i)(t-i-1)f^{u-i}h^{i-l}e^{t-i-1} \\
&= \sum_{i=0}^{u}\sum_{l=0}^{i}\sum_{k=0}^{i-l}\gamma_{u,i,l}\binom{i-l}{k}(-2)^k f^{u+1-i}h^{i-l-k}e^{t-i} \\
&\quad +\sum_{i=0}^{u}\sum_{l=0}^{i}\gamma_{u,i,l}(t-i)f^{u-i}h^{i-l+1}e^{t-i-1} \\
&\quad -\sum_{i=0}^{u}\sum_{l=0}^{i}\gamma_{u,i,l}(t-i)(t-i-1)f^{u-i}h^{i-l}e^{t-i-1}.
\end{aligned}$$

Now we selectively reindex and rename indices as necessary. To begin, we separate the three distinct summations in the above:

$$\sum_{i=0}^{u}\sum_{l=0}^{i}\sum_{k=0}^{i-l}\gamma_{u,i,l}\binom{i-l}{k}(-2)^k f^{u+1-i}h^{i-l-k}e^{t-i}, \tag{18}$$

$$\sum_{i=0}^{u}\sum_{l=0}^{i}\gamma_{u,i,l}(t-i)f^{u-i}h^{i-l+1}e^{t-i-1}, \tag{19}$$

$$\sum_{i=0}^{u}\sum_{l=0}^{i}\gamma_{u,i,l}(t-i)(t-i-1)f^{u-i}h^{i-l}e^{t-i-1}. \tag{20}$$

Concentrating on (19), we reindex by setting $j=i+1$, in which case we have

$$\sum_{j=1}^{u+1}\sum_{l=0}^{j-1}\gamma_{u,j-1,l}(t-j+1)f^{u+1-j}h^{j-l}e^{t-j}.$$

By our conventions on the constant $\gamma_{u,i,l}$ we may extend the second summation to $l=j$ without changing the value. Additionally, if $j=0$, we have $\gamma_{u,-1,l}$, which is also zero by convention. Thus we may rewrite (19) as

$$\sum_{j=0}^{u+1}\sum_{l=0}^{j}\gamma_{u,j-1,l}(t-j+1)f^{u+1-j}h^{j-l}e^{t-j},$$

which matches the required form.

Turning our attention to (20), we make the same reindexing with the substitution $j=i+1$, thereby yielding

$$\sum_{j=1}^{u+1}\sum_{l=0}^{j-1}\gamma_{u,j-1,l}(t-j+1)(t-j)f^{u+1-j}h^{j-1-l}e^{t-j}.$$

Replacing l by $s=l+1$ we have

$$\sum_{j=1}^{u+1}\sum_{s=1}^{j}\gamma_{u,j-1,s-1}(t-j+1)(t-j)f^{u+1-j}h^{j-s}e^{t-j}$$

and once again, if we extend the sums to $s=0$ and $j=0$ respectively, the resulting scalars are zero. Thus (20) becomes

$$\sum_{j=0}^{u+1}\sum_{s=0}^{j}\gamma_{u,j-1,s-1}(t-j+1)(t-j)f^{u+1-j}h^{j-s}e^{t-j}.$$

We rename the indices using the original index names in the second and third parts and consolidate to yield

$$\sum_{i=0}^{u+1}\sum_{l=0}^{j}\big(\gamma_{u,i-1,l}(t-i+1)-\gamma_{u,i-1,l-1}(t-i)(t-i-1)\big)f^{u+1-i}h^{i-l}e^{t-i}. \tag{21}$$

Finally turning our attention to (18), to achieve the desired form we must reindex the sum using the substitution $r=l+k$. Starting with binomial coefficent, note that

$$\binom{i-l}{k}=\binom{i-l}{i-l-k}=\binom{i-l}{i-r}$$

and that the power $(-2)^k$ is the same as $(-2)^{r-l}$. In terms of the summation bounds, if l ranges between 0 and i and k ranges between 0 and $i-l$, this is equivalent to l ranging between 0 and i and $r=l+k$ ranging between l and i. Thus (18) becomes

$$\sum_{i=0}^{u+1}\sum_{l=0}^{i}\sum_{r=l}^{i}\gamma_{u,i,l}\binom{i-l}{i-r}(-2)^{r-l}f^{u+1-i}h^{i-r}e^{t-i}. \tag{22}$$

Once again, we have extended the summation bounds using the conventions on $\gamma_{u,i,l}$. This gives the PBW basis elements in terms of the indices i and r for this part. Renaming l in (22) by s and r by l and consolidating with (21) we have that $e^t f^{u+1}$ can be expressed as

$$\begin{aligned} e^t f^{u+1} = &\sum_{i=0}^{u+1}\sum_{s=0}^{i}\sum_{l=s}^{i}\gamma_{u,i,s}\binom{i-s}{i-r}(-2)^{l-s}f^{u+1-i}h^{i-l}e^{t-i} \\ &+\sum_{i=0}^{u+1}\sum_{l=0}^{j}\big(\gamma_{u,i-1,l}(t-i+1)-\gamma_{u,i-1,l-1}(t-i)(t-i-1)\big)f^{u+1-i}h^{i-l}e^{t-i}. \end{aligned} \tag{23}$$

This gives $e^t f^{u+1}$ as a linear combination of terms of the form $f^{u+1-i}h^{i-l}e^{t-i}$ as required.

It is easier to establish the recursive formula if we use (18) in its original formulation. Fixing a and b with $b \le a$, the coefficient of $f^{u+a-1}h^{a-b}e^{t-a}$ in (18) is

$$\sum_{l+k=0}^{b}\binom{a-l}{k}(-2)^k\gamma_{u,a,l},$$

while the remaining term comes from setting $i=a$ and $l=b$ in (23) giving

$$\begin{aligned} \gamma_{u+1,a,b} = &\left(\sum_{l+k=0}^{b}\binom{a-l}{k}(-2)^k\gamma_{u,a,l}\right) \\ &+\gamma_{u,a-1,b}(t-a+1)-\gamma_{u,a-1,b-1}(t-a)(t-a-1), \end{aligned} \tag{24}$$

completing the proof. □

We can rewrite the sum in the last equation a little more explicitly as

$$\sum_{l=0}^{a}\sum_{k=b-l}^{a-l}\binom{a-l}{k}(-2)^k\gamma_{u,a,l}$$

if need be.

5. Main result

We are now ready to prove the main result, namely expressing $(f^r h^s e^t)(f^u h^v e^w)$ in terms of the PBW basis using the preceding results. Given that

$$e^t f^u = \sum_{i=0}^{u}\sum_{l=0}^{i}\gamma_{u,i,l}f^{u-i}h^{i-l}e^{t-i},$$

we have

$$
\begin{aligned}
&(f^r h^s e^t)(f^u h^v e^w)\\
&= f^r h^s (e^t f^u) h^v e^w\\
&= f^r h^s \left(\sum_{i=0}^{u} \sum_{l=0}^{i} \gamma_{u,i,l} f^{u-i} h^{i-l} e^{t-i} \right) h^v e^w\\
&= f^r h^s \left(\sum_{i=0}^{u} \sum_{l=0}^{i} \gamma_{u,i,l} f^{u-i} h^{i-l} (e^{t-i} h^v) \right) e^w\\
&= f^r h^s \left(\sum_{i=0}^{u} \sum_{l=0}^{i} \gamma_{u,i,l} f^{u-i} h^{i-l} (h-2(t-i))^v e^{t-i} \right) e^w\\
&= f^r \sum_{i=0}^{u} \sum_{l=0}^{i} \gamma_{u,i,l} (h^s f^{u-i}) h^{i-l} (h-2(t-i))^v e^{w+t-i}\\
&= f^r \sum_{i=0}^{u} \sum_{l=0}^{i} \gamma_{u,i,l} f^{u-i} (h-2(u-i))^s h^{i-l} (h-2(t-i))^v e^{w+t-i}\\
&= \sum_{i=0}^{u} \sum_{l=0}^{i} \gamma_{u,i,l} f^{r+u-i} (h-2(u-i))^s h^{i-l} (h-2(t-i))^v e^{w+t-i}\\
&= \sum_{i=0}^{u} \sum_{l=0}^{i} \sum_{j=0}^{s} \sum_{k=0}^{v} \gamma_{u,i,l} \binom{s}{j} \binom{v}{k} (-2)^{j+k} (u-i)^j (t-i)^k f^{r+u-i} h^{s+v+i-(j+k+l)} e^{t+w-i}.
\end{aligned}
$$

The above, together with the recursion given by (24) allows the computation of the structure constants.

References

[Carter 2005] R. W. Carter, *Lie algebras of finite and affine type*, Cambridge Studies in Advanced Mathematics **96**, Cambridge University Press, 2005. MR Zbl

[Erdmann and Wildon 2006] K. Erdmann and M. J. Wildon, *Introduction to Lie algebras*, Springer, 2006. MR Zbl

[Humphreys 1972] J. E. Humphreys, *Introduction to Lie algebras and representation theory*, Graduate Texts in Mathematics **9**, Springer, 1972. MR Zbl

Received: 2013-06-03 Revised: 2013-11-13 Accepted: 2020-08-06

agourley@masonlive.gmu.edu *Department of Statistics, George Mason University, Fairfax, VA, United States*

christopher.kennedy@cnu.edu *Department of Mathematics, Christopher Newport University, Newport News, VA, United States*

INVOLVE 13:4 (2020)
https://doi.org/10.2140/involve.2020.13.551

Conjecture $\mathcal{O}$ holds for some horospherical varieties of Picard rank 1

Lela Bones, Garrett Fowler, Lisa Schneider and Ryan M. Shifler

(Communicated by Kenneth S. Berenhaut)

Property $\mathcal{O}$ for an arbitrary complex, Fano manifold X is a statement about the eigenvalues of the linear operator obtained from the quantum multiplication of the anticanonical class of X. Conjecture $\mathcal{O}$ is a conjecture that property $\mathcal{O}$ holds for any Fano variety. Pasquier classified the smooth nonhomogeneous horospherical varieties of Picard rank 1 into five classes. Conjecture $\mathcal{O}$ has already been shown to hold for the odd symplectic Grassmannians, which is one of these classes. We will show that conjecture $\mathcal{O}$ holds for two more classes and an example in a third class of Pasquier's list. Perron–Frobenius theory reduces our proofs to be graph-theoretic in nature.

1. Introduction

The purpose of this paper is to prove that conjecture $\mathcal{O}$ holds for some horospherical varieties of Picard rank 1. We recall the precise statement of conjecture $\mathcal{O}$ for varieties of Picard rank 1, following [Galkin et al. 2016, Section 3]. Let F be a Fano variety, let $K := K_F$ be the canonical line bundle of F, let F_D be a fundamental divisor of F, and let

$$c_1(F) := c_1(-K) \in H^2(F)$$

be the anticanonical class. The Fano index of F is r, where r is the greatest integer such that $K_F \cong -rF_D$. The small quantum cohomology ring $(QH^*(F), \star)$ is a graded algebra over $\mathbb{Z}[q]$, where q is the quantum parameter. We define the small quantum cohomology in Section 2.1. Consider the specialization

$$H^\bullet(F) := QH^*(F)|_{q=1}$$

at $q = 1$. The quantum multiplication by the first Chern class $c_1(F)$ induces an endomorphism $\hat{c}_1$ of the finite-dimensional vector space $H^\bullet(F)$:

$$y \in H^\bullet(F) \mapsto \hat{c}_1(y) := (c_1(F) \star y)|_{q=1}.$$

MSC2010: primary 14N35; secondary 14N15, 15B48.
Keywords: quantum cohomology, horospherical, conjecture $\mathcal{O}$.

Set $\delta_0 := \max\{|\delta| : \delta \text{ is an eigenvalue of } \hat{c}_1\}$. Then property $\mathcal{O}$ states the following:

(1) The real number δ_0 is an eigenvalue of $\hat{c}_1$ of multiplicity 1.

(2) If δ is any eigenvalue of $\hat{c}_1$ with $|\delta| = \delta_0$, then $\delta = \delta_0\gamma$ for some r-th root of unity $\gamma \in \mathbb{C}$, where r is the Fano index of F.

Property $\mathcal{O}$ was conjectured to hold for any Fano, complex manifold F in [Galkin et al. 2016]. If a Fano, complex, manifold has property $\mathcal{O}$ then we say that the space satisfies conjecture $\mathcal{O}$. Conjecture $\mathcal{O}$ underlies gamma conjectures I and II of Galkin, Golyshev, and Iritani. The gamma conjectures refine earlier conjectures by Dubrovin on Frobenius manifolds and mirror symmetry. Conjecture $\mathcal{O}$ has already been proved for the homogeneous G/P case in [Cheong and Li 2017], the odd symplectic Grassmannians in [Li et al. 2019], del Pezzo surfaces in [Hu et al. 2019], and projective complete intersections in [Ke 2018]. The Perron–Frobenius theory of nonnegative matrices reduces the proofs that conjecture $\mathcal{O}$ holds for the homogeneous and the odd symplectic Grassmannian cases to be a graph-theoretic check. This is because conjecture $\mathcal{O}$ is largely reminiscent of Perron–Frobenius theory. In this manuscript we will use the same graph-theoretic approach to prove that conjecture $\mathcal{O}$ holds for some smooth horospherical varieties of Picard rank 1.

Next we recall the definition of a horospherical variety following [Gonzales et al. 2018]. Let G be a complex reductive group. A G-variety is a reduced scheme of finite type over the field of complex numbers $\mathbb{C}$, equipped with an algebraic action of G. Let B be a Borel subgroup of G. A G-variety X is called spherical if X has a dense B-orbit. Let X be a G-spherical variety and let H be the stabilizer of a point in the dense G-orbit in X. The variety X is called *horospherical* if H contains a conjugate of the maximal unipotent subgroup of G contained in the Borel subgroup B.

Smooth horospherical varieties of Picard rank 1 were classified in [Pasquier 2009]. These varieties are either homogeneous or can be constructed in a uniform way via a triple $(\text{Type}(G), \omega_Y, \omega_Z)$ of representation-theoretic data, where $\text{Type}(G)$ is the semisimple Lie type of the reductive group G and ω_Y, ω_Z are fundamental weights. See [Pasquier 2009, Section 1.3] for details. Pasquier classified the possible triples into five classes:

(1) $(B_n, \omega_{n-1}, \omega_n)$ with $n \geq 3$.

(2) $(B_3, \omega_1, \omega_3)$.

(3) $(C_n, \omega_m, \omega_{m-1})$ with $n \geq 2$ and $m \in [2, n]$ (the odd symplectic Grassmannians).

(4) $(F_4, \omega_2, \omega_3)$.

(5) $(G_2, \omega_1, \omega_2)$.

Proposition 3.6 [Pasquier 2006] showed the triples in the above list are Fano varieties. We are now able to state the main theorem:

Theorem 1. *If F belongs to the classes* (1) *for* $n = 3$, (2), (3), *and* (5) *of Pasquier's list, then conjecture $\mathcal{O}$ holds for F.*

2. Preliminaries

2.1. ***Quantum cohomology.*** The small quantum cohomology is defined as follows. Let $(\alpha_i)_i$ be a basis of $H^*(F)$, the classical cohomology ring, and let $(\alpha_i^\vee)_i$ be the dual basis for the Poincaré pairing. The multiplication is given by

$$\alpha_i \star \alpha_j = \sum_{d\geq 0,k} c_{i,j}^{k,d} q^d \alpha_k,$$

where $c_{i,j}^{k,d}$ are the 3-point, genus-0, Gromov–Witten invariants corresponding to the classes α_i, α_j, and $\alpha_k^\vee$. We will make use of the quantum Chevalley formula, which is the multiplication of a hyperplane class h with another class α_j. Theorem 0.0.3 of [Gonzales et al. 2018] implies that if F belongs to the classes (1) for $n = 3$, (2), or (5) of Pasquier's list, then there is an explicit quantum Chevalley formula. The explicit quantum Chevalley formula is the key ingredient used to prove property $\mathcal{O}$ holds.

2.2. ***Sufficient criterion for property $\mathcal{O}$ to hold.*** We recall the notion of the (oriented) quantum Bruhat graph of a Fano variety F. The vertices of this graph are the basis elements

$$\alpha_i \in H^\bullet(F) := QH^*(F)|_{q=1}.$$

There is an oriented edge $\alpha_i \to \alpha_j$ if the class α_j appears with positive coefficient (where we consider $q > 0$) in the quantum Chevalley multiplication $h \star \alpha_i$ for some hyperplane class h. Using the Perron–Frobenius theory of nonnegative matrices, conjecture $\mathcal{O}$ reduces to a graph-theoretic check of the quantum Bruhat graph. The techniques involving Perron–Frobenius theory used by Li, Mihalcea, and Shifler [Li et al. 2019] and Cheong and Li [2017] imply the following lemma:

Lemma 2. *Suppose the following conditions hold for a Fano variety F:*

(1) *The matrix representation of $\hat{c}_1$ is nonnegative.*

(2) *The quantum Bruhat graph of F is strongly connected.*

(3) *There exists a cycle of length r, the Fano index, in the quantum Bruhat graph of F.*

Then property $\mathcal{O}$ holds for F. We say the matrix representation of $\hat{c}_1$ is nonnegative if all of the entries are nonnegative.

We refer the reader to [Minc 1988, Section 4.3] for further details on Perron–Frobenius theory.

3. Checking property $\mathcal{O}$ holds

Let X be a horospherical variety. We will simplify our notation where the basis of $H^\bullet(X)$ is $\{1, h, \alpha_i\}_{i\in I}$ for some finite index set I. Observe by [Gonzales et al. 2018] that the anticanonical classes are

$$c_1(X) = \begin{cases} 5h & \text{when } X \text{ is case (1) for } n = 3, \\ 7h & \text{when } X \text{ is case (2)}, \\ 4h & \text{when } X \text{ is case (5)}, \end{cases}$$

and the Fano indices are

$$r = \begin{cases} 5 & \text{when } X \text{ is case (1) for } n = 3, \\ 7 & \text{when } X \text{ is case (2)}, \\ 4 & \text{when } X \text{ is case (5)}. \end{cases}$$

The endomorphism $\hat{c}_1$ acting on the basis elements of $H^\bullet(X)$ is determined by the Chevalley formula in the following way:

$$\begin{aligned} \hat{c}_1(\alpha_i) &= 5(h \star \alpha_i)|_{q=1} \quad \text{when } X \text{ is case (1) for } n = 3, \\ \hat{c}_1(\alpha_i) &= 7(h \star \alpha_i)|_{q=1} \quad \text{when } X \text{ is case (2), and} \\ \hat{c}_1(\alpha_i) &= 4(h \star \alpha_i)|_{q=1} \quad \text{when } X \text{ is case (5)}. \end{aligned}$$

Each of the following three subsections will show that conjecture $\mathcal{O}$ holds for case (1) for $n = 3$, case (2), and case (5) of Pasquier's list, respectively. In each subsection we will reformulate the quantum Chevalley formulas stated in [Gonzales et al. 2018], present the quantum Bruhat graph, and argue that each condition of Lemma 2 is satisfied. For each case, we have kept the same format of the equations presented by [Gonzales et al. 2018] with our prescribed basis for ease of identification for the reader. For example, line 3 in Proposition 4.3 of that paper is

$$h * \sigma'_{u_2} = \sigma'_{u_3} + \sigma'_{u'_3} \quad \text{and} \quad h * \sigma'_{u'_2} = 2\sigma'_{u'_3} + \tau_{v_0}.$$

In Proposition 3 below we identify this line with

$$\hat{c}_1(\alpha_1) = 5\alpha_3 + 5\alpha_4 \quad \text{and} \quad \hat{c}_1(\alpha_2) = 10\alpha_3 + 5\alpha_5.$$

3.1. ***Case* (1) *for* $n = 3$.** We will reformulate the quantum Chevalley formula stated in [Gonzales et al. 2018] using the basis $\{1, h, \alpha_1, \alpha_2, \ldots, \alpha_{18}\}$.

Proposition 3. *The following equalities hold by* [Gonzales et al. 2018, Proposition 4.3]:

(1) $\hat{c}_1(1) = 5h$.

(2) $\hat{c}_1(h) = 10\alpha_1 + 5\alpha_2$.

(3) $\hat{c}_1(\alpha_1) = 5\alpha_3 + 5\alpha_4$ *and* $\hat{c}_1(\alpha_2) = 10\alpha_3 + 5\alpha_5$.

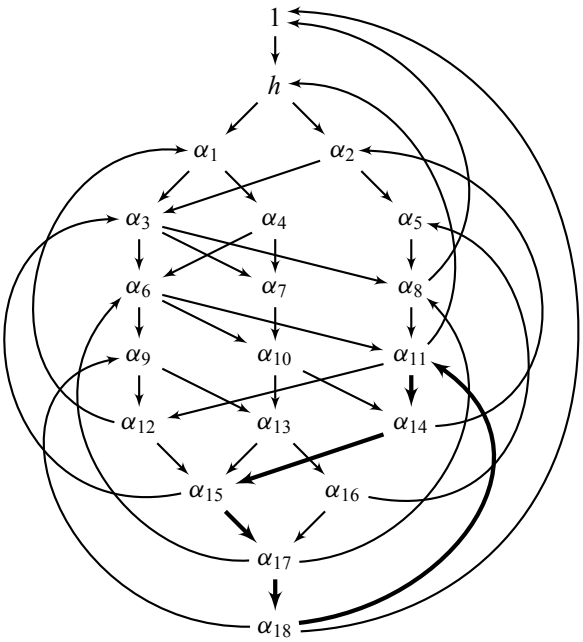


Figure 1. The quantum Bruhat graph of the Fano variety X in case (1) for $n = 3$.

(4) $\hat{c}_1(\alpha_3) = 10\alpha_6 + 5\alpha_7 + 5\alpha_8$, $\hat{c}_1(\alpha_4) = 5\alpha_6 + 10\alpha_7$, *and* $\hat{c}_1(\alpha_5) = 5\alpha_8$.

(5) $\hat{c}_1(\alpha_6) = 10\alpha_9 + 5\alpha_{10} + 5\alpha_{11}$, $\hat{c}_1(\alpha_7) = 5\alpha_{10}$, *and* $\hat{c}_1(\alpha_8) = 5\alpha_{11} + 5 \cdot 1$.

(6) $\hat{c}_1(\alpha_9) = 5\alpha_{12} + 5\alpha_{13}$, $\hat{c}_1(\alpha_{10}) = 10\alpha_{13} + 5\alpha_{14}$, $\hat{c}_1(\alpha_{11}) = 5\alpha_{12} + 5\alpha_{14} + 5h$.

(7) $\hat{c}_1(\alpha_{12}) = 5\alpha_{15} + 5\alpha_1$, $\hat{c}_1(\alpha_{13}) = 5\alpha_{15} + 5\alpha_{16}$, *and* $\hat{c}_1(\alpha_{14}) = 5\alpha_{15} + 5\alpha_2$.

(8) $\hat{c}_1(\alpha_{15}) = 5\alpha_{17} + 5\alpha_3$ *and* $\hat{c}_1(\alpha_{16}) = 5\alpha_{17} + 5\alpha_5$.

(9) $\hat{c}_1(\alpha_{17}) = 5\alpha_{18} + 5\alpha_6 + 5\alpha_8$.

(10) $\hat{c}_1(\alpha_{18}) = 5\alpha_9 + 5\alpha_{11} + 10 \cdot 1$.

The quantum Bruhat graph is shown in Figure 1. The bold edges indicate a cycle of length $r = 5$, the Fano index.

Lemma 4. *Property $\mathcal{O}$ holds when X is case* (1) *with $n = 3$ of Pasquier's list.*

Proof. The coefficients that appear in the equations in Proposition 3 are the entries of the matrix representation of $\hat{c}_1$. Therefore, the matrix representation of $\hat{c}_1$ is nonnegative. The quantum Bruhat graph is strongly connected by Figure 1, and the cycle $\alpha_{18}\alpha_{11}\alpha_{14}\alpha_{15}\alpha_{17}\alpha_{18}$ has length $r = 5$. □

3.2. ***Case* (2).** Again, we reformulate the quantum Chevalley formula from [Gonzales et al. 2018] using the basis $\{1, h, \alpha_1, \alpha_2, \ldots, \alpha_{12}\}$.

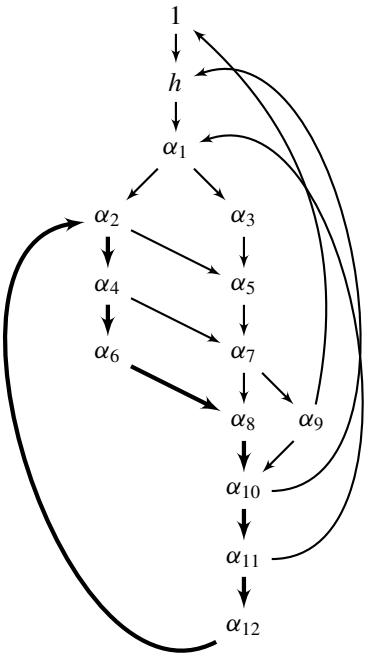


Figure 2. The quantum Bruhat graph for case (2).

Proposition 5. *The following equalities hold by* [Gonzales et al. 2018, Proposition 4.4]:

(1) $\hat{c}_1(1) = 7h$.

(2) $\hat{c}_1(h) = 7\alpha_1$.

(3) $\hat{c}_1(\alpha_1) = 14\alpha_2 + 7\alpha_3$.

(4) $\hat{c}_1(\alpha_2) = 7\alpha_4 + 7\alpha_5$ *and* $\hat{c}_1(\alpha_3) = 7\alpha_5$.

(5) $\hat{c}_1(\alpha_4) = 7\alpha_6 + 7\alpha_7$ *and* $\hat{c}_1(\alpha_5) = 7\alpha_7$.

(6) $\hat{c}_1(\alpha_6) = 7\alpha_8$ *and* $\hat{c}_1(\alpha_7) = 7\alpha_8 + 7\alpha_9$.

(7) $\hat{c}_1(\alpha_8) = 7\alpha_{10}$ *and* $\hat{c}_1(\alpha_9) = 7\alpha_{10} + 7 \cdot 1$.

(8) $\hat{c}_1(\alpha_{10}) = 7\alpha_{11} + 7h$.

(9) $\hat{c}_1(\alpha_{11}) = 7\alpha_{12} + 7\alpha_1$.

(10) $\hat{c}_1(\alpha_{12}) = 7\alpha_2$.

The quantum Bruhat graph is shown in Figure 2.

Lemma 6. *Property $\mathcal{O}$ holds when X is case* (2) *of Pasquier's list.*

Proof. The coefficients that appear in the equations in Proposition 5 are the entries of the matrix representation of $\hat{c}_1$. Therefore, the matrix representation of $\hat{c}_1$ is

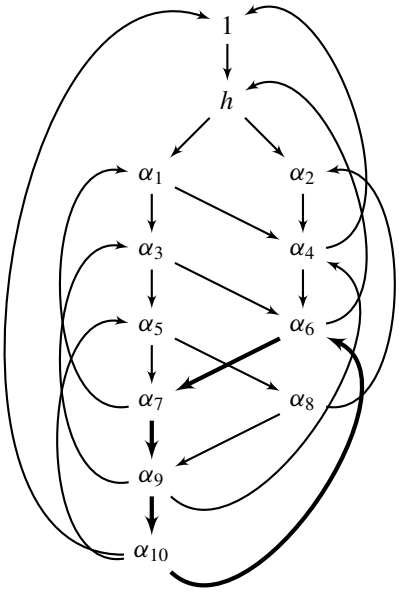


Figure 3. The quantum Bruhat graph for case (5).

nonnegative. The quantum Bruhat graph is strongly connected by Figure 2, and the cycle $\alpha_{12}\alpha_2\alpha_4\alpha_6\alpha_8\alpha_{10}\alpha_{11}\alpha_{12}$ has length $r = 7$. □

3.3. ***Case* (5).** Again, we reformulate the quantum Chevalley formula from [Gonzales et al. 2018] using the basis $\{1, h, \alpha_1, \alpha_2, \ldots, \alpha_{10}\}$.

Proposition 7. *The following equalities hold by* [Gonzales et al. 2018, Proposition 4.6]:

(1) $\hat{c}_1(1) = 4h$.

(2) $\hat{c}_1(h) = 12\alpha_1 + 4\alpha_2$.

(3) $\hat{c}_1(\alpha_1) = 8\alpha_3 + 4\alpha_4$ *and* $\hat{c}_1(\alpha_2) = 4\alpha_4$.

(4) $\hat{c}_1(\alpha_3) = 12\alpha_5 + 4\alpha_6$ *and* $\hat{c}_1(\alpha_4) = 4\alpha_6 + 4 \cdot 1$.

(5) $\hat{c}_1(\alpha_5) = 4\alpha_7 + 4\alpha_8$ *and* $\hat{c}_1(\alpha_6) = 8\alpha_7 + 4h$.

(6) $\hat{c}_1(\alpha_7) = 4\alpha_9 + 4\alpha_1$ *and* $\hat{c}_1(\alpha_8) = 4\alpha_9 + 4\alpha_2$.

(7) $\hat{c}_1(\alpha_9) = 4\alpha_{10} + 4\alpha_3 + 4\alpha_4$.

(8) $\hat{c}_1(\alpha_{10}) = 4\alpha_5 + 4\alpha_6 + 8 \cdot 1$.

The associated quantum Bruhat graph is shown in Figure 3.

Lemma 8. *Property $\mathcal{O}$ holds when X is case* (5) *of Pasquier's list.*

Proof. The coefficients that appear in the equations in Proposition 7 are the entries of the matrix representation of $\hat{c}_1$. Therefore, the matrix representation of $\hat{c}_1$ is nonnegative. The quantum Bruhat graph is strongly connected by Figure 3, and the cycle $\alpha_{10}\alpha_6\alpha_7\alpha_9\alpha_{10}$ has length $r = 4$. □

Theorem 1 follows from Lemmas 4, 6, 8, and the previously mentioned work done by Li, Mihalcea, Shifler [Li et al. 2019] for the odd symplectic Grassmannian case.

Acknowledgement

We would like to thank the anonymous referees for their useful comments and suggestions to improve the presentation of this paper.

References

[Cheong and Li 2017] D. Cheong and C. Li, "On the conjecture $\mathcal{O}$ of GGI for G/P", *Adv. Math.* **306** (2017), 704–721. MR Zbl

[Galkin et al. 2016] S. Galkin, V. Golyshev, and H. Iritani, "Gamma classes and quantum cohomology of Fano manifolds: gamma conjectures", *Duke Math. J.* **165**:11 (2016), 2005–2077. MR Zbl

[Gonzales et al. 2018] R. Gonzales, C. Pech, N. Perrin, and A. Samokhin, "Geometry of horospherical varieties of Picard rank one", preprint, 2018. arXiv

[Hu et al. 2019] J. Hu, H.-Z. Ke, C. Li, and T. Yang, "Gamma conjecture I for del Pezzo surfaces", preprint, 2019. arXiv

[Ke 2018] H.-Z. Ke, "On Conjecture $\mathcal{O}$ for projective complete intersections", preprint, 2018. arXiv

[Li et al. 2019] C. Li, L. C. Mihalcea, and R. M. Shifler, "Conjecture $\mathcal{O}$ holds for the odd symplectic Grassmannian", *Bull. Lond. Math. Soc.* **51**:4 (2019), 705–714. MR Zbl

[Minc 1988] H. Minc, *Nonnegative matrices*, John Wiley & Sons, New York, 1988. MR Zbl

[Pasquier 2006] B. Pasquier, *Variétés horosphèriques de Fano*, Ph.D. thesis, Université Joseph Fourier, 2006, available at https://tel.archives-ouvertes.fr/tel-00111912/document.

[Pasquier 2009] B. Pasquier, "On some smooth projective two-orbit varieties with Picard number 1", *Math. Ann.* **344**:4 (2009), 963–987. MR Zbl

Received: 2018-07-10 Revised: 2019-07-12 Accepted: 2020-07-25

lbones1@gulls.salisbury.edu *Department of Mathematics, Salisbury University, Salisbury, MD, United States*

gfowler2@gulls.salisbury.edu *Department of Mathematics, Salisbury University, Salisbury, MD, United States*

lmschneider@salisbury.edu *Department of Mathematics, Salisbury University, Salisbury, MD, United States*

rmshifler@salisbury.edu *Department of Mathematics, Salisbury University, Salisbury, MD, United States*

INVOLVE 13:4 (2020)
https://doi.org/10.2140/involve.2020.13.559

Condensed Ricci curvature of complete and strongly regular graphs

Vincent Bonini, Conor Carroll, Uyen Dinh,
Sydney Dye, Joshua Frederick and Erin Pearse

(Communicated by Kenneth S. Berenhaut)

We study a modified notion of Ollivier's coarse Ricci curvature on graphs introduced by Lin, Lu, and Yau. We establish a rigidity theorem for complete graphs that shows a connected finite simple graph is complete if and only if the Ricci curvature is strictly greater than 1. We then derive explicit Ricci curvature formulas for strongly regular graphs in terms of the graph parameters and the size of a maximal matching in the core neighborhood. As a consequence we are able to derive exact Ricci curvature formulas for strongly regular graphs of girths 4 and 5 using elementary means. An example is provided that shows there is no exact formula for the Ricci curvature for strongly regular graphs of girth 3 that is purely in terms of graph parameters.

1. Introduction

Lott and Villani [2009] and Sturm [2006] discovered a relationship between optimal transport and Ricci curvature on smooth Riemannian manifolds and they pursued a construction of a synthetic notion of Ricci curvature that could be defined independently of differentiable structures. Ollivier [2009; 2010] later introduced a notion of coarse Ricci curvature for Markov chains on metric spaces which has a particularly accessible formulation on graphs. Much work has been done on coarse Ricci curvature on graphs; see [Bauer et al. 2012; Lin et al. 2011; 2014; Jost and Liu 2014; Smith 2014; Bhattacharya and Mukherjee 2015; Bourne et al. 2018; Münch and Wojciechowski 2019].

Ollivier's coarse Ricci curvature on graphs is defined in terms of transport distance of probability measures (see Section 2). We study a modified notion of Ollivier's coarse Ricci curvature on graphs introduced by Lin, Lu, and Yau [Lin et al. 2011]. In this paper we refer to the modified Ricci curvature of [Lin et al. 2011] as the condensed Ricci curvature. Our first result concerning the condensed Ricci

MSC2010: primary 52C99, 53B99; secondary 05C10, 05C81, 05C99.
Keywords: coarse Ricci curvature, strongly regular graphs.

curvature exploits the relationship between Ricci curvature and the eigenvalues of the spectral graph Laplacian to establish a rigidity theorem for complete graphs. This result is stated without proof in [Lin et al. 2011, Example 1].

Theorem 1.1. *A connected finite simple graph $G = (V, E)$ is complete if and only if the condensed Ricci curvature satisfies $\Bbbk(x, y) > 1$ for all edges $xy \in E$. In this case, $\Bbbk(x, y) = n/(n-1)$ for all vertices $x, y \in V$.*

We then turn our attention to the derivation of exact formulas for the condensed Ricci curvature on strongly regular graphs. Explicit formulas and curvature bounds for various forms of Ricci curvature have been established for large classes of graphs; see [Lin et al. 2011; 2014; Jost and Liu 2014; Smith 2014; Bhattacharya and Mukherjee 2015; Cushing et al. 2020]. In particular, Bakry–Émery Ricci curvature functions are studied in [Cushing et al. 2020] and explicit formulas for the Bakry–Émery Ricci curvature are derived for Cayley graphs and strongly regular graphs of girths 4 and 5. This work served as our inspiration to establish explicit formulas for the condensed Ricci curvature on strongly regular graphs using elementary means.

Let $G = (V, E)$ be a strongly regular graph with parameters (n, d, α, β). Here $n = |V|$ is the number of vertices, d is the uniform degree of the vertices, $\alpha \geq 0$ is the number of common neighbors for adjacent vertices, and $\beta \geq 1$ is the number of common neighbors for nonadjacent vertices. Given an edge $xy \in E$, let N_x denote the set of vertices that are adjacent to x, not including y or any vertices that are adjacent to y. Similarly, let N_y denote the set of vertices that are adjacent to y, not including x or any vertices that are adjacent to x (see Section 2 for more details). We compute the following curvature formulas solely from properties of maximum matchings and the regularity properties of strongly regular graphs.

Theorem 1.2. *Suppose $G = (V, E)$ is a strongly regular graph with parameters (n, d, α, β), and $xy \in E$ with maximum matching $\mathcal{M}$ of size m between N_x and N_y. Then the condensed Ricci curvature satisfies*

$$\Bbbk(x, y) = \frac{\alpha + 2}{d} - \frac{|N_x| - m}{d}. \tag{1-1}$$

Remark 1.3. We would like to point out that Theorem 1.2 holds more generally for regular graphs of diameter 2. In this case one takes α to be the number of neighbors common to both x and y. For simplicity in our presentation, we focus on strongly regular graphs.

The explicit formula $\Bbbk(x, y) = (\alpha + 2)/d$ was previously established in [Smith 2014, Theorem 6.3] for general graphs with perfect matchings between N_x and N_y. In the setting of strongly regular graphs Theorem 1.2 is thus an extension of the results of [Smith 2014] to the case of maximum matchings. These results give

insight into the importance of understanding matchings between the neighbor sets N_x and N_y in the derivation of exact Ricci curvature formulas. In particular, we are able to exploit this idea to derive the explicit formulas for the Ricci curvature of strongly regular graphs of girths 4 and 5 using elementary means.

For strongly regular graphs $G = (V, E)$ of girth 5, it follows that adjacent vertices share $\alpha = 0$ common neighbors. Moreover, for a given edge $xy \in E$, there are no edges between the neighbor sets N_x and N_y. Then, as a simple consequence of Theorem 1.2, we have the following Ricci curvature formula.

Theorem 1.4. *Suppose $G = (V, E)$ is a strongly regular graph with parameters (n, d, α, β). If the girth of G is 5, then the condensed Ricci curvature satisfies*

$$\Bbbk(x, y) = \frac{3}{d} - 1$$

for all $xy \in E$.

From Theorem 1.4 it is easy to see that the condensed Ricci curvature along any edge of the 5-cycle C_5 is $\Bbbk = \frac{1}{2}$, while the condensed Ricci curvature along any edge of the Peterson graph is $\Bbbk = 0$. The fact that these are the only nonnegatively curved strongly regular graphs of girth 5 then follows from Theorem 1.4 and the classification of Moore graphs of diameter 2 due to [Hoffman and Singleton 1960].

Corollary 1.5. *The 5-cycle and the Petersen graph are the only strongly regular graphs of girth 5 with nonnegative condensed Ricci curvature along edges.*

For strongly regular graphs $G = (V, E)$ of girth 4, it follows that adjacent vertices share $\alpha = 0$ common neighbors. For a given edge $xy \in E$, we appeal to Hall's theorem and the pigeonhole principle to establish a perfect matching between the neighbor sets N_x and N_y. Then, as a consequence of Theorem 1.2, we have the following Ricci curvature formula.

Theorem 1.6. *Suppose $G = (V, E)$ is a strongly regular graph with parameters (n, d, α, β). If the girth of G is 4, then the condensed Ricci curvature satisfies*

$$\Bbbk(x, y) = \frac{2}{d}$$

for all $xy \in E$. In particular, all strongly regular graphs of girth 4 have positive condensed Ricci curvature.

For strongly regular graphs with girth 3, it turns out that there does not exist an exact formula for the condensed Ricci curvature solely in terms of the graph parameters. This is due to the fact that there are nonisomorphic girth-3 strongly regular graphs with the same graph parameters but different condensed Ricci curvatures. For example, the 4×4 rook's graph and the Shrikhande graph both have parameters $(16, 6, 2, 2)$ but their condensed Ricci curvatures are $\Bbbk = \frac{2}{3}$ and $\Bbbk = \frac{1}{3}$, respectively. This is due to the fact that for a given edge xy there is a

perfect matching between neighbor sets N_x and N_y for the rook's graph but the maximum matching between N_x and N_y is size $m = 1$ for the Shrikhande graph. For comparison, we would like to point out that the 4×4 rook's graph and the Shrikhande graph were given as an example in [Cushing et al. 2020] to show that Bakry–Émery Ricci curvature functions are not uniquely determined by strongly regular graph parameters.

On the other hand, strongly regular conference graphs exhibit many symmetry properties that we believe always lead to perfect matchings between neighbor sets N_x and N_y. We plan to address conference graphs in a forthcoming paper. For now we make the following conjecture for conference graphs, which can be realized as a direct consequence of Theorem 1.2 provided there are always perfect matchings between N_x and N_y.

Conjecture 1.7. *Let $G = (V, E)$ be a strongly regular conference graph with parameters $(4\beta + 1, 2\beta, \beta - 1, \beta)$, with $\beta \geq 2$. Then the condensed Ricci curvature satisfies*

$$\Bbbk(x, y) = \frac{1}{2} + \frac{1}{2\beta}$$

for all $xy \in E$. In particular, all strongly regular conference graphs have positive condensed Ricci curvature strictly greater than $\frac{1}{2}$.

This paper is organized as follows: In Section 2 we formally define transport distance between probability measures on graphs and the condensed Ricci curvature. We briefly discuss our choice of notation and terminology and we state some key results from [Bourne et al. 2018; Bhattacharya and Mukherjee 2015] that simplify the computation of condensed Ricci curvature. In Section 3 we introduce the spectral graph Laplacian and prove Theorem 1.1. In Section 4 we focus on strongly regular graphs and establish our exact curvature formulas.

2. Transport distance and condensed Ricci curvature on graphs

To motivate Ollivier's formulation of coarse Ricci curvature we turn to the setting of smooth Riemannian manifolds. Suppose that x_1 and x_2 are points on an N-dimensional Riemannian manifold M and that m_i is the uniform measure on the respective geodesic ball $B(x_i, \varepsilon)$ of radius ε centered at x_i. Let $\delta = d(x_1, x_2)$ denote the geodesic distance between x_1 and x_2. If $B(x_1, \varepsilon)$ is parallel transported to $B(x_2, \varepsilon)$ along a geodesic from x_1 to x_2 in the direction of tangent vector v at x_1, then the transport distance $W(m_1, m_2)$ between the measures and the Ricci curvature on M satisfy the relation [Ollivier 2009]

$$W(m_1, m_2) = \delta\left(1 - \frac{\varepsilon^2}{2(N+2)}\operatorname{Ric}(v, v) + O(\varepsilon^3 + \varepsilon^2\delta)\right) \quad \text{as } \varepsilon, \delta \to 0.$$

As noted in [Ollivier 2010], this relation allows one to interpret Ricci curvature on Riemannian manifolds as a measure of the average distance traveled between points of geodesic balls versus their centers under parallel transport. Using this relationship to guide intuition, Ollivier defined the Ricci curvature of Markov chains on metric spaces as

$$\mathrm{Ric}(x_1, x_2) = 1 - \frac{W(m_1, m_2)}{d(x_1, x_2)},$$

where $W(m_1, m_2)$ denotes the general transport or Wasserstein distance between measures m_i based at x_i and $d(x_1, x_2)$ is the metric distance between x_1 and x_2. This definition provides a synthetic notion of Ricci curvature of Markov chains on metric spaces and its geometric validation comes in the form of analogues of the classical theorems of Bonnet–Myers, Lichnerowicz, and Gromov–Lévy [Ollivier 2009; 2010].

Ollivier's coarse Ricci curvature on graphs is defined in terms of transport distance of probability measures and standard graph distance [Ollivier 2009; 2010]. Let $G = (V, E)$ be a nontrivial, locally finite, undirected, connected, simple graph with standard graph shortest-path distance function

$$\rho : V \times V \to \mathbb{N} \cup \{0\}.$$

A *probability measure* or *probability mass distribution* on $G = (V, E)$ is a real-valued function $\mu : V \to [0, 1]$ such that

$$\sum_{v \in V} \mu(v) = 1.$$

Given probability measures μ and ν on G, a *coupling* or *transport plan* between μ and ν is a probability measure $\xi : V \times V \to [0, 1]$ such that

$$\sum_{w \in V} \xi(v, w) = \mu \quad \text{and} \quad \sum_{v \in V} \xi(v, w) = \nu.$$

The terminology transport plan has economic origins and comes from the optimal transport problem, where one interprets a coupling as a plan for transporting mass or goods from one distribution center to another.

The L^1-Wasserstein or transport distance is a metric on the space of probability measures on G that quantifies the optimal *transport cost* between probability measures on G.

Definition 2.1. The L^1-*Wasserstein* or *transport distance* between probability measures μ and ν on a graph $G = (V, E)$ is given by

$$W(\mu, \nu) = \inf_{\xi \in \chi(\mu, \nu)} \sum_{x \in V} \sum_{y \in V} \rho(x, y) \xi(x, y),$$

where $\chi(\mu, \nu)$ is the set of all transport plans between μ and ν.

We consider a modified notion of Ollivier's coarse Ricci curvature on graphs introduced by Lin, Lu, and Yau [Lin et al. 2011] that is based on probability measures of the form

$$m_x^\varepsilon(y) = \begin{cases} 1-\varepsilon & y = x, \\ \varepsilon/\deg(x) & y \in \Gamma(x), \\ 0 & \text{otherwise.} \end{cases} \tag{2-1}$$

For $x, y \in V$ the ε-Ollivier Ricci curvature is then defined for $0 \le \varepsilon \le 1$ by

$$\kappa_\varepsilon(x, y) = 1 - \frac{W(m_x^\varepsilon, m_y^\varepsilon)}{\rho(x, y)}$$

and the modified Ollivier Ricci curvature is then defined for $x, y \in V$ by

$$\Bbbk(x, y) = \lim_{\varepsilon \to 0} \frac{\kappa_\varepsilon(x, y)}{\varepsilon}.$$

We refer to the modified Ricci curvature as the condensed Ricci curvature. Our definitions differ slightly from those of [Lin et al. 2011] in that we use $1-\varepsilon$ in place of α as in [Smith 2014; Münch and Wojciechowski 2019]. The reason for this choice in notation and terminology is that we interpret the probability measure (2-1) as encoding an ε-active random walk that becomes stationary as $\varepsilon \to 0$. One can view this phenomena as a concentration or condensation of measure. This formulation also more closely resembles the notion of coarse Ricci curvature of continuous-time Markov chains given by the derivative

$$\kappa(x, y) = -\frac{d}{dt} \frac{W(m_x^t, m_y^t)}{\rho(x, y)}.$$

Computing Ricci curvature on graphs requires solving the optimal transport problem. By definition a given transport plan between measures m_x^ε and m_x^ε gives an upper bound for the transport distance $W(m_x^\varepsilon, m_y^\varepsilon)$ and therefore a lower bound for condensed Ricci curvature. While an optimal transport plan may be intuitive, realizing that it achieves the transport distance directly from the definition is technically challenging as it requires solving a linear programming problem. Fortunately, there is a dual formulation of this optimization problem in terms of 1-Lipschitz functions that one can use to provide lower bounds to transport distance and consequently upper bounds for condensed Ricci curvature.

Theorem 2.2 (Kantorovich duality theorem). *The L^1-Wasserstein or transport distance between probability measures μ and ν on a graph $G = (V, E)$ is given by*

$$W(\mu, \nu) = \sup_{f \in \mathrm{Lip}(1)} \sum_{z \in V} f(z)(\mu(z) - \nu(z)),$$

where

$$\mathrm{Lip}(1) = \{f : V \to \mathbb{R} \mid |f(x) - f(y)| \leq \rho(x, y) \text{ for all } x, y \in V\}$$

is the space of 1*-Lipschitz functions on* G.

Due to the Kantorovich duality theorem we refer to 1-Lipschitz functions on a graph $G = (V, E)$ as Kantorovich potential functions. With the Kantorovich duality theorem in hand one can compute exact formulas for Ricci curvature by specifying a transport plan and a Kantorovich potential that provide bounds that pinch the condensed Ricci curvature and force equality. Another useful fact for computing condensed Ricci curvature due to [Bourne et al. 2018] is that the Ricci activeness function $\varepsilon \mapsto \Bbbk_\varepsilon$ is piecewise linear with two linear parts on regular graphs of degree d and the ε-Ollivier Ricci curvature differs from the condensed Ricci curvature only by scale for all values $0 \leq \varepsilon \leq d/(d+1)$. More precisely, on a regular graph $G = (V, E)$ of degree d, for any $x, y \in V$,

$$\kappa_\varepsilon(x, y) = \varepsilon \Bbbk(x, y) \quad \text{for } 0 \leq \varepsilon \leq \frac{d}{d+1}.$$

For simplicity, in our computations we take $\varepsilon = \frac{1}{2}$ and compute the condensed Ricci curvature as

$$\Bbbk(x, y) = 2\kappa_{1/2}(x, y).$$

In Riemannian geometry, Ricci curvature is a local quantity so in the context of graphs it seems natural to restrict one's attention to computing Ricci curvature along edges. With our conventions, this restriction allows for further simplification in computing Ricci curvature due to [Bhattacharya and Mukherjee 2015]. For vertices $x \in V$ let

$$\Gamma(x) = \{y \in V \mid \rho(x, y) = 1\} = \{y \in V \mid xy \in E\}$$

denote the neighbor set of x. Given an edge $xy \in E$ we denote the set of common neighbors or *triangle set* of x and y by

$$\nabla_{xy} = \Gamma(x) \cap \Gamma(y).$$

Then we further decompose the 1-step neighborhoods of x and y into disjoint unions

$$\Gamma(x) = N_x \cup \nabla_{xy} \cup \{y\} \quad \text{and} \quad \Gamma(y) = N_y \cup \nabla_{xy} \cup \{x\}$$

where

$$N_x = \Gamma(x) \setminus (\nabla_{xy} \cup \{y\}) \quad \text{and} \quad N_y = \Gamma(y) \setminus (\nabla_{xy} \cup \{x\}).$$

We denote the set of common 2-step neighbors or *pentagon set* of x and y by

$$P_{xy} = \{z \in V \mid \rho(x, z) = 2 \text{ and } \rho(y, z) = 2\}.$$

As in [Bhattacharya and Mukherjee 2015], we refer to the disjoint union of vertices

$$\mathcal{N}_{xy} = \{x\} \cup \{y\} \cup \nabla_{xy} \cup N_x \cup N_y \cup P_{xy} \tag{2-2}$$

as the *core neighborhood* of an edge $xy \in E$. It is immediate that the core neighborhood represents a partition of all vertices with distance less than or equal to 2 from x and y. Moreover, the reduction lemma [Bhattacharya and Mukherjee 2015, Lemma 2.3] shows that when computing the condensed Ricci curvature of an edge $xy \in E$, it is sufficient to consider only the induced subgraph of G lying within the core neighborhood $\mathcal{N}_{xy}$.

3. A rigidity theorem for complete graphs

In this section we establish a rigidity theorem for complete graphs that shows that a complete graph is the only graph with condensed Ricci curvature strictly greater than 1. Given a connected finite simple graph $G = (V, E)$ with n vertices $v_1, \ldots, v_n$, the adjacency matrix A of G is a symmetric $n \times n$ matrix with entries

$$\begin{cases} a_{ij} = 1 & \text{if } v_i v_j \in E, \\ a_{ij} = 0 & \text{otherwise.} \end{cases}$$

Let d_i denote the degree of vertex v_i and let $D = \operatorname{diag}(d_1, \ldots, d_n)$ be the $n \times n$ diagonal matrix of degrees. The spectral graph Laplacian is then defined as

$$L = I - D^{-1/2} A D^{-1/2}.$$

With these conventions it follows that the eigenvalues of the spectral graph Laplacian are nondecreasing with

$$0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1},$$

where the multiplicity of the 0 eigenvalue is the number of connected components of G. Hence, a finite graph is connected if and only if $\lambda_1 > \lambda_0 = 0$. Similar to the spectrum of the metric Laplacian on Riemannian manifolds, the spectrum of the spectral Laplacian on graphs encodes geometric and structural properties of a graph.

In the context of Riemannian geometry the Lichnerowicz theorem [1958] is a widely celebrated theorem that provides a lower bound for the first nonzero eigenvalue of the metric Laplacian when the Ricci curvature has a strict positive lower bound. More precisely, if (M^n, g) is an n-dimensional compact Riemannian manifold with

$$\operatorname{Ric}_g \geq (n-1)kg$$

for some positive constant $k > 0$, then the first positive eigenvalue of the metric Laplacian $\lambda_1 \geq nk$. The graph analogue of the Lichnerowicz theorem similarly shows that a positive lower bound on condensed Ricci curvature serves as a lower bound on the first nonzero eigenvalue of the spectral graph Laplacian.

Theorem 3.1 [Lin et al. 2011, Theorem 4.2]. *Suppose $G = (V, E)$ is a connected finite simple graph with condensed Ricci curvature*

$$\Bbbk(x, y) \geq \kappa_0 > 0$$

for some positive constant κ_0 and all edges $xy \in E$. Then the first nonzero eigenvalue of the spectral graph Laplacian satisfies $\lambda_1 \geq \kappa_0$.

In order to establish a rigidity theorem for complete graphs we first record some well-known results on the first nonzero eigenvalue of the spectral graph Laplacian.

Theorem 3.2 [Chung 1997, Lemma 1.7]. *Let $G = (V, E)$ be a simple graph on n vertices. Then*

$$\lambda_1 \leq \frac{n}{n-1}$$

with equality if and only if G is complete. Moreover, complete graphs are the only graphs with $\lambda_1 > 1$.

We now prove the following rigidity theorem through direct computation and the aid of Theorems 3.1 and 3.2.

Theorem 3.3. *A connected finite simple graph $G = (V, E)$ is complete if and only if the condensed Ricci curvature satisfies $\Bbbk(x, y) > 1$ for all edges $xy \in E$.*

Proof. Let $G = (V, E)$ be the complete graph on n vertices and let $xy \in E$. Then the triangle set ∇_{xy} is equal to $V \setminus \{x, y\}$ since G is complete. Let m_x and m_y denote $\varepsilon = \frac{1}{2}$-probability measures based at x and y, respectively, defined as in (2-1). Consider the transport plan $\xi : V \times V \to [0, 1]$ between m_x and m_y given by

$$\xi(v, w) = \begin{cases} \frac{1}{2} - \frac{1}{2(n-1)}, & v = x,\ w = y, \\ \frac{1}{2(n-1)}, & v = w, \\ 0, & \text{otherwise}, \end{cases}$$

which transports excess mass directly from x to y and leaves the remainder of the distribution fixed. Then by Definition 2.1 the transport distance satisfies

$$W(m_x, m_y) \leq \sum_{v \in V} \sum_{w \in V} \rho(v, w)\xi(v, w) = \xi(x, y) = \frac{1}{2} - \frac{1}{2(n-1)}$$

and therefore the condensed Ricci curvature satisfies

$$\Bbbk(x, y) = 2\kappa_{1/2}(x, y) \geq 2\left(1 - \left(\frac{1}{2} - \frac{1}{2(n-1)}\right)\right) = 1 + \frac{1}{n-1} > 1. \tag{3-1}$$

Conversely, let $G = (V, E)$ be a connected finite simple graph with condensed Ricci curvature $\Bbbk(x, y) > 1$ for all edges $xy \in E$. Then by the graph Lichnerowicz theorem (Theorem 3.1) of [Lin et al. 2011], it follows $\lambda_1 > 1$ and therefore G is complete by Theorem 3.2. □

Remark 3.4. Using the 1-Lipschitz function $f : V \to \mathbb{R}$ defined by

$$f(v) = \begin{cases} 1, & v = x, \\ 0, & \text{otherwise}, \end{cases}$$

it follows from the Kantorovich duality theorem (Theorem 2.2) that the transport distance satisfies

$$W(m_x, m_y) \geq \sum_{v \in V} f(v)(m_x(v) - m_y(v)) = m_x(x) - m_y(x) = \frac{1}{2} - \frac{1}{2(n-1)}$$

and therefore the condensed Ricci curvature satisfies

$$\Bbbk(x, y) = 2\kappa_{1/2}(x, y) \leq 2\left(1 - \left(\frac{1}{2} - \frac{1}{2(n-1)}\right)\right) = 1 + \frac{1}{n-1}. \tag{3-2}$$

Hence, from (3-1) and (3-2) we see that the condensed Ricci curvature of the complete graph on n vertices is

$$\Bbbk(x, y) = 1 + \frac{1}{n-1} = \frac{n}{n-1}$$

for all edges $xy \in E$ as stated in [Lin et al. 2011, Example 1].

4. Strongly regular graphs

In this section we derive an explicit formula for the condensed Ricci curvature of strongly regular graphs $G = (V, E)$ in terms of the graph parameters and the size of a maximum matching in the core neighborhood. Recall that a simple, undirected, finite graph $G = (V, E)$ is said to be a strongly regular graph with parameters (n, d, α, β) if G has n vertices of constant degree d where any two adjacent vertices share $\alpha \geq 0$ common neighbors and any two nonadjacent vertices share $\beta \geq 1$ common neighbors. Given an edge $xy \in E$, we consider the core neighborhood decomposition (2-2) given by

$$\mathcal{N}_{xy} = \{x\} \cup \{y\} \cup \nabla_{xy} \cup N_x \cup N_y \cup P_{xy}$$

that was introduced in [Bhattacharya and Mukherjee 2015] and where the neighbor sets satisfy

$$N_x = \Gamma(x) \setminus (\nabla_{xy} \cup \{y\}) \quad \text{and} \quad N_y = \Gamma(x) \setminus (\nabla_{xy} \cup \{y\}).$$

Since strongly regular graphs have diameter 2, it follows that $V = \mathcal{N}_{xy}$. Hence, for strongly regular graphs the induced subgraph lying within the core neighborhood of any edge accounts for the entire graph G. Moreover, from strong regularity we see $|\nabla_{xy}| = \alpha$ so that

$$|N_x| = |N_y| = d - \alpha - 1 \quad \text{and} \quad |P_{xy}| = n - 2d - \alpha.$$

Our strategy in the derivation of an explicit formula for the condensed Ricci curvature of strongly regular graphs reformulates the problem of finding candidates for an optimal transport plan and an optimal Kantorovich potential into a maximum matching problem.

Definition 4.1. A *matching* on a finite undirected simple graph $G = (V, E)$ is a subset of edges $\mathcal{M} \subseteq E$ such that no vertex in V is incident to more than one edge in $\mathcal{M}$. The size of a matching $\mathcal{M}$ on G is the total number of edges in $\mathcal{M}$. A matching $\mathcal{M}$ on G is said to be a *maximum matching* if it contains the largest possible number of edges. A maximum matching between disjoint subsets of vertices $S, T \subset V$ is said to be a *perfect matching* if every vertex in S and T is incident to exactly one edge of the matching.

Let $G = (V, E)$ be a strongly regular graph with edge $xy \in E$. Given a maximum matching $\mathcal{M}$ between the neighbor sets N_x and N_y, we denote the vertices that are matched in N_x and N_y by M_x and M_y, respectively. If $\mathcal{M}$ is not a perfect matching, we denote the remaining unmatched vertices in N_x and N_y by $U_x = N_x \setminus M_x$ and $U_y = N_y \setminus M_y$, respectively. With this decomposition of N_x and N_y into matched and unmatched vertices, it is fairly straightforward to develop a candidate for an optimal transport plan. However, developing a candidate for an optimal Kantorovich potential requires a deeper understanding of the incidence relations between edges in a given maximum matching and the remaining edges between N_x and N_y.

Definition 4.2. Let $\mathcal{M}$ be a matching on an undirected simple graph $G = (V, E)$. An *alternating path* P in G is a path $P = v_0 v_1 \cdots v_k$ such that the initial vertex v_0 is unmatched and $v_i v_{i+1} \in \mathcal{M}$ if and only if $v_{i-1} v_i \notin \mathcal{M}$ for all $0 < i < k$. An *augmenting path* P in G is an alternating path $P = v_0 v_1 \cdots v_k$ such that the terminal vertex v_k is also unmatched.

We refer to a single vertex as a trivial path. Hence, with this convention any unmatched vertex is by itself a trivial augmenting path. One of the main tools in our construction of a candidate for an optimal Kantorovich potential is Berge's lemma [1957].

Lemma 4.3 (Berge's lemma). *A matching $\mathcal{M}$ on an undirected simple graph $G = (V, E)$ is a maximum matching if and only if G contains no nontrivial augmenting paths.*

Given a matching $\mathcal{M}$ on a graph $G = (V, E)$ and a subset of vertices $U \subseteq V$, we denote the subset of matched vertices in U by $\mathcal{M}(U)$. For a subset of vertices $S \subseteq V$ we denote the subset of alternating paths with initial vertex in S by A_S. For subsets of vertices $S, T \subseteq V$ we denote the set of vertices in T that lie along an alternating path initiated in S by $A_S(T)$. We use the contrapositive of the following lemma to show that certain vertices in the neighbor sets N_x and N_y have no edges

between them. Ultimately, this allows us to ensure our candidate for an optimal Kantorovich potential is 1-Lipschitz.

Lemma 4.4. *Let $H = (V, E)$ be a finite, undirected, bipartite graph with parts S, T. Let $\mathcal{M}$ be a maximum matching on H and suppose $v \in A_S(S)$ and $w \in T$. If $vw \in E$, then $w \in A_S(T)$.*

Proof. First note that since H is bipartite, it follows that an alternating path initiated in S can only terminate in an unmatched vertex in T. Therefore, if $v \in A_S(S)$ is an unmatched vertex, then v must be the initial vertex of an alternating path. But then w must be a matched vertex so that $w \in A_S(T)$ since otherwise $vw \in E$ is an augmenting path, contradicting Berge's lemma as $\mathcal{M}$ is a maximum matching. Now assume that v is a matched vertex and that $P = v_0 v_1 \cdots v_i v$ is an alternating path initiated at $v_0 \neq v \in S$ that terminates at $v \in S$. Then since $v_0 \in S$ is an unmatched vertex and H is bipartite, it follows that $v_i v \in \mathcal{M}$ since P is an alternating path. But then $vw \notin \mathcal{M}$ since otherwise v is incident to more than one edge in $\mathcal{M}$. But then the concatenation $P' = v_0 v_1 \cdots v_i vw$ is an alternating path so that $w \in A_S(T)$. □

We use this final lemma to compute a lower bound for condensed Ricci curvature from our Kantorovich potential in terms of the number of vertices in N_x and N_y along alternating paths initiated in N_y.

Lemma 4.5. *Let $H = (V, E)$ be a finite, undirected, bipartite graph with parts S, T. Suppose $\mathcal{M}$ is a maximum matching of size m on H. Then*

$$|A_S(S)| = |A_S(T)| + |S| - m.$$

Proof. Let $v \in A_S(T)$ and suppose $P = v_0 v_1 \cdots v_i v$ is an alternating path initiated at $v_0 \in S$ that terminates at $v \in T$. Then since $v_0 \in S$ is an unmatched vertex and H is bipartite, it follows that $v_i v \notin \mathcal{M}$ since P is an alternating path. If v is an unmatched vertex, then P is an augmenting path contrary to Berge's lemma as $\mathcal{M}$ is a maximum matching. Hence, v must be a matched vertex and therefore since matched vertices are incident to a single edge in the matching, it follows that there exits a unique vertex $w \in S$ such that $vw \in \mathcal{M}$. But then the concatenation $P' = v_0 v_1 \cdots vw$ is an alternating path initiated in S that terminates at the matched vertex $w \in \mathcal{M}(S)$. Hence, for each $v \in A_S(T)$ there is a unique $w \in A_S(S) \cap \mathcal{M}(S)$ with $vw \in \mathcal{M}$.

On the other hand, given $w \in A_S(S) \cap \mathcal{M}(S)$ suppose $P = w_0 w_1 \cdots w_i w$ is an alternating path initiated in S that terminates at $w \in S$. Then since w is a matched vertex by assumption, it is clear that P is a nontrivial path and that $w_i \in T$ is the unique matched vertex with $w_i w \in \mathcal{M}$. But then $P' = w_0 w_1 \cdots w_i$ is an alternating path initiated in S that terminates at $w_i \in T$ so that $w_i \in A_S(T)$. Hence, for each $w \in A_S(S) \cap \mathcal{M}(S)$ there is a unique $v \in A_S(T)$ with $vw \in \mathcal{M}$.

Thus, since $A_S(T) \subseteq \mathcal{M}(T)$, we have a bijection between the subsets $A_S(T)$ and $A_S(S) \cap \mathcal{M}(S)$ and therefore

$$|A_S(T)| = |A_S(S) \cap \mathcal{M}(S)|.$$

Noting $S \setminus \mathcal{M}(S) \subseteq A_S(S)$ since under our conventions all unmatched vertices in S are trivial alternating paths initiated in S, it follows

$$\begin{aligned} |A_S(T)| &= |A_S(S) \cap \mathcal{M}(S)| = |A_S(S) \setminus (S \setminus \mathcal{M}(S))| \\ &= |A_S(S)| - |S \setminus \mathcal{M}(S)| = |A_S(S)| - (|S| - m). \end{aligned}$$

Hence,

$$|A_S(S)| = |A_S(T)| + |S| - m$$

as desired. □

We are now in a position to derive an explicit formula for condensed Ricci curvature of an edge $xy \in E$ in terms of the graph parameters and the size of a maximum matching between the neighbor sets N_x and N_y.

Theorem 4.6. *Suppose $G = (V, E)$ is a strongly regular graph with parameters (n, d, α, β), and $xy \in E$ with maximum matching $\mathcal{M}$ of size m between N_x and N_y. Then the condensed Ricci curvature satisfies*

$$\Bbbk(x, y) = \frac{\alpha + 2}{d} - \frac{|N_x| - m}{d}.$$

Proof. Let $xy \in E$ and consider the core neighborhood decomposition of G as in (2-2). Let H denote the induced bipartite subgraph consisting of all edges in E between vertices in N_x and N_y. Suppose that $\mathcal{M}$ is a maximum matching on H of size $|\mathcal{M}| = m$. Assuming $\mathcal{M}$ is not a perfect matching, as above we denote the matched and unmatched vertices in N_x and N_y by M_x, M_y and U_x, U_y, respectively. Note that $|U_x| = |U_y|$ since $|N_x| = |N_y|$ and $|M_x| = |M_y|$. For $i = 1, \ldots, |N_x| - m$, let $x_i^u \in N_x$ and $y_i^u \in N_y$ denote the unmatched vertices in H and for $j = 1, \ldots, m$, let $x_j^m \in N_x$ and $y_j^m \in N_y$ denote the matched vertices in H. Up to reordering of indices we may assume $x_i^m y_i^m \in \mathcal{M}$ for $i = 1, \ldots, m$.

Clearly, there are no edges in E shared by two unmatched vertices since $\mathcal{M}$ is a maximum matching so $x_i^u y_j^u \notin E$ for all $1 \leq i, j \leq |N_x| - m$. But then $\rho(x_i^u, y_i^u) = 2$ in G since x_i^u and y_i^u are not adjacent and G has diameter 2. Hence, for $1 \leq i \leq |N_x| - m$, we may pair unmatched vertices x_i^u with y_i^u along 2-step paths in G. For simplicity in notation write $x_i^u y_i^u \in \mathcal{M}^2$ to denote unmatched vertices $x_i^u \in U_x$ and $y_i^u \in U_y$ that have been paired along a 2-step path. Together these matchings/pairings induce a transport plan that moves excess mass from x to y and that moves mass from $x_i^m \in M_x$, $x_i^u \in U_x$ directly to their matched/paired neighbors $y_i^m \in M_y$, $y_i^u \in U_y$ along 1-step and 2-step paths, respectively. More precisely,

consider the transport plan defined by

$$\xi(v,w)=\begin{cases}\frac{1}{2}-\frac{1}{2d}, & v=x,\ w=y,\\ \frac{1}{2d}, & v=w\in\nabla_{xy}\cup\{x\}\cup\{y\},\\ \frac{1}{2d}, & vw\in\mathcal{M},\\ \frac{1}{2d}, & vw\in\mathcal{M}^2,\\ 0, & \text{otherwise.}\end{cases}\tag{4-1}$$

Noting that $|N_x|=|N_y|=d-\alpha-1$ it follows that the transport distance satisfies

$$\begin{aligned}W(m_x,m_y)&\le\sum_{v\in V}\sum_{w\in V}\rho(v,w)\xi(v,w)\\&=\left(\frac{1}{2}-\frac{1}{2d}\right)+m\cdot\frac{1}{2d}+(|N_x|-m)\cdot 2\cdot\frac{1}{2d}\\&=\frac{1}{2}+\frac{1}{2d}((d-\alpha-2)+(|N_x|-m)).\end{aligned}$$

Thus, the condensed Ricci curvature satisfies

$$\begin{aligned}\Bbbk(x,y)&=2\kappa_{1/2}(x,y)\\&\ge 2\left(1-\left(\frac{1}{2}+\frac{1}{2d}((d-\alpha-2)+(|N_x|-m))\right)\right)\\&=\frac{\alpha+2}{d}-\frac{|N_x|-m}{d}.\end{aligned}$$

Next we define a Kantorovich potential function $f:V\to\mathbb{R}$ with respect to the matching $\mathcal{M}$ by

$$f(z)=\begin{cases}1, & z=x,\\ 1, & z\in N_x\setminus A_{N_y}(N_x),\\ -1, & z\in A_{N_y}(N_y),\\ 0, & \text{otherwise,}\end{cases}\tag{4-2}$$

where $A_{N_y}(N_x)$ and $A_{N_y}(N_y)$ denote the vertices in N_x and N_y, respectively, that lie along alternating paths initiated in N_y. Taking $T=N_x$ and $S=N_y$ in Lemma 4.4, it follows that there are no edges between $N_x\setminus A_{N_y}(N_x)$ and $A_{N_y}(N_y)$. Hence, f is a 1-Lipschitz function on V.

Let $a_x=|A_{N_y}(N_x)|$ and $a_y=|A_{N_y}(N_y)|$. Then

$$|N_x\setminus A_{N_y}(N_x)|=|N_x|-a_x=d-\alpha-1-a_x$$

and by Lemma 4.5

$$a_y=|A_{N_y}(N_y)|=|A_{N_y}(N_x)|+|N_y|-m=a_x+|N_x|-m.$$

Therefore, by the Kantorovich duality theorem, we find that the transport distance satisfies

$$\begin{aligned} W(m_x, m_y) &\geq \sum_{z\in V} f(z)(m_x(z)-m_y(z)) = \left(\frac{1}{2}-\frac{1}{2d}\right)+(d-\alpha-1-a_x)\cdot\frac{1}{2d}+a_y\cdot\frac{1}{2d} \\ &= \left(\frac{1}{2}-\frac{1}{2d}\right)+(d-\alpha-1-a_x)\frac{1}{2d}+(a_x+|N_x|-m)\frac{1}{2d} \\ &= \frac{1}{2}+\frac{1}{2d}((d-\alpha-2)+(|N_x|-m)). \end{aligned}$$

Similar to the calculation (4-2), we have

$$\Bbbk(x, y) = 2\kappa_{1/2}(x, y) \leq \frac{\alpha+2}{d} - \frac{|N_x|-m}{d}. \tag{4-3}$$

Hence, from the inequalities (4-2) and (4-3), it follows

$$\Bbbk(x, y) = \frac{\alpha+2}{d} - \frac{|N_x|-m}{d}.$$

Now if $\mathcal{M}$ is a perfect matching between N_x and N_y, then the size of $\mathcal{M}$ is by definition $m = |N_x| = |N_y| = d-\alpha-1$. Moreover, since all vertices in N_x and N_y are matched by some edge in $\mathcal{M}$, it follows that there are no alternating paths initiated in N_x or N_y in the induced bipartite subgraph between vertices in N_x and N_y. But then $A_{N_y}(N_x) = \varnothing$ and $A_{N_y}(N_y) = \varnothing$. Taking all of this into account, we see that the same transport plan (4-1) coupled with the Kantorovich potential (4-2) yields the equality

$$\Bbbk(x, y) = \frac{\alpha+2}{d}$$

in the case of a perfect matching between N_x and N_y. □

Strongly regular graphs of girths 4 and 5. In this subsection we exploit the adjacency properties of strongly regular graphs $G = (V, E)$ to derive explicit formulas for the condensed Ricci curvature of strongly regular graphs of girths 4 and 5. Let $G = (V, E)$ be a strongly regular graph with parameters (n, d, α, β). It is well known [Cameron and van Lint 1975] that if G has girth 5, then $\alpha = 0$ and $\beta = 1$, while a strongly regular graph G with girth 4 has $\alpha = 0$ and $\beta \geq 2$. The following theorem is an immediate consequence of Theorem 4.6 and the fact that the girth of a graph is its minimal cycle length.

Theorem 4.7. *Suppose $G = (V, E)$ is a strongly regular graph with parameters (n, d, α, β). If the girth of G is 5, then the condensed Ricci curvature satisfies*

$$\Bbbk(x, y) = \frac{3}{d} - 1$$

for all edges $xy \in E$.

Proof. Let $xy \in E$ and consider the core neighborhood decomposition of G as in (2-2). Let $x_0 \in N_x$ and suppose that $x_0 y_0 \in E$ for some $y_0 \in N_y$. Then $x\,x_0\,y_0\,y\,x$ is a cycle of length 4, contradicting the fact that G has girth 5. Thus, there are no edges between N_x and N_y so the maximum matching between N_x and N_y has size $m = 0$. Therefore, from Theorem 4.6 with $\alpha = 0$ and $m = 0$, we find that the condensed Ricci curvature satisfies

$$\Bbbk(x, y) = \frac{3}{d} - 1. \qquad \square$$

Remark 4.8. The transportation plan (4-1) and Kantorovich potential (4-2) can be used to compute this curvature formula directly. Since G has girth 5, there are no edges between vertices in N_x and N_y so $\mathcal{M} = \varnothing$ and each vertex in N_x can be paired to a vertex in N_y through a 2-step path. Moreover, if we consider the induced bipartite subgraph H between N_x and N_y, the collection of alternating paths in H that start in N_y consists solely of trivial paths so $A_{N_y}(N_y) = N_y$ and $A_{N_y}(N_x) = \varnothing$.

In the girth 4 case we appeal to Hall's theorem, which can be used to give necessary and sufficient conditions for perfect matchings on bipartite graphs. Let $H = (V, E)$ be a finite, bipartite graph with parts S, T. For any subset $W \subseteq S$, we define

$$\Gamma_T(W) = \{v \in T \mid vw \in E \text{ for some } w \in W\}$$

to be the collection of all vertices $v \in T$ sharing an edge with some vertex in W. More simply put, $\Gamma_T(W)$ is the neighbor set of W in T.

Theorem 4.9 (Hall's theorem). *Let $H = (V, E)$ be a finite, undirected, bipartite graph with parts S, T. Then there is a matching that covers S if and only if*

$$|\Gamma_T(W)| \geq |W|$$

for every subset of vertices $W \subseteq S$.

With Hall's theorem in hand we are able to establish a perfect matching between the neighbor sets N_x and N_y for strongly regular graphs with girth 4 as a consequence of the pigeonhole principle.

Theorem 4.10. *Suppose $G = (V, E)$ is a strongly regular graph with parameters (n, d, α, β). If the girth of G is 4, then the condensed Ricci curvature satisfies*

$$\Bbbk(x, y) = \frac{2}{d}$$

for all edges $xy \in E$.

Proof. Let $xy \in E$ and consider the core neighborhood decomposition of G as in (2-2). Then since $\alpha = 0$, it follows that G is triangle free. Suppose that $x_i \in N_x$. Then since x_i is nonadjacent to y, it follows x_i and y have $\beta \geq 2$ common neighbors.

Clearly, one such neighbor is x so x_i has $\beta - 1 \geq 1$ neighbors in N_y since $\nabla_{xy} = \varnothing$. Similarly, each vertex $y_j \in N_y$ has $\beta - 1 \geq 1$ neighbors in N_x.

Now consider the induced bipartite subgraph $H = G[N_x \cup N_y]$ consisting of all edges in E between vertices in N_x and N_y. For sake of contradiction suppose that $X \subseteq N_x$ with $|X| = q > |\Gamma_{N_y}(X)| = k$. Since each vertex $x_i \in X \subseteq N_x$ has $\beta - 1 \geq 1$ neighbors in N_y, it follows that there are $(\beta - 1)q$ edges between X and $\Gamma_{N_y}(X)$. Therefore, by the generalized pigeonhole principle there exists some $w \in \Gamma_{N_y}(X)$ with

$$|\Gamma(w)| \geq \left\lceil \frac{(\beta - 1)q}{k} \right\rceil > \left\lceil \frac{(\beta - 1)q}{q} \right\rceil = \beta - 1$$

since $q > k$. This contradicts the fact that every vertex in N_y has exactly $\beta - 1$ neighbors in N_x. Thus, $|X| \leq |\Gamma_{N_y}(X)|$ for all subsets $X \subset N_x$ and therefore since $|N_x| = |N_y|$ it follows that there is a perfect matching in H by Hall's theorem. Noting that $\alpha = 0$, it follows from Theorem 4.6 with $\alpha = 0$ and $m = |N_x|$ that the condensed Ricci curvature satisfies

$$\Bbbk(x, y) = \frac{2}{d}. \qquad \square$$

Acknowledgements

The authors were partially supported by NSF FURST grant DMS-1620552, the Cal Poly Frost Fund, CSU Fresno, and Cal Poly, San Luis Obispo. The authors would like to thank CSU Fresno for their hospitality during their summer participation in the FURST program.

References

[Bauer et al. 2012] F. Bauer, J. Jost, and S. Liu, "Ollivier–Ricci curvature and the spectrum of the normalized graph Laplace operator", *Math. Res. Lett.* **19**:6 (2012), 1185–1205. MR Zbl

[Berge 1957] C. Berge, "Two theorems in graph theory", *Proc. Nat. Acad. Sci. U.S.A.* **43** (1957), 842–844. MR Zbl

[Bhattacharya and Mukherjee 2015] B. B. Bhattacharya and S. Mukherjee, "Exact and asymptotic results on coarse Ricci curvature of graphs", *Discrete Math.* **338**:1 (2015), 23–42. MR Zbl

[Bourne et al. 2018] D. P. Bourne, D. Cushing, S. Liu, F. Münch, and N. Peyerimhoff, "Ollivier–Ricci idleness functions of graphs", *SIAM J. Discrete Math.* **32**:2 (2018), 1408–1424. MR Zbl

[Cameron and van Lint 1975] P. J. Cameron and J. H. van Lint, *Graph theory, coding theory and block designs*, London Math. Soc. Lecture Note Series **19**, Cambridge University Press, 1975. MR Zbl

[Chung 1997] F. R. K. Chung, *Spectral graph theory*, CBMS Regional Conference Series in Mathematics **92**, American Math. Soc., Providence, RI, 1997. MR Zbl

[Cushing et al. 2020] D. Cushing, S. Liu, and N. Peyerimhoff, "Bakry–Émery curvature functions on graphs", *Canad. J. Math.* **72**:1 (2020), 89–143. MR Zbl

[Hoffman and Singleton 1960] A. J. Hoffman and R. R. Singleton, "On Moore graphs with diameters 2 and 3", *IBM J. Res. Develop.* **4** (1960), 497–504. MR Zbl

[Jost and Liu 2014] J. Jost and S. Liu, "Ollivier's Ricci curvature, local clustering and curvature-dimension inequalities on graphs", *Discrete Comput. Geom.* **51**:2 (2014), 300–322. MR Zbl

[Lichnerowicz 1958] A. Lichnerowicz, *Géométrie des groupes de transformations*, Travaux et Recherches Mathématiques, III, Dunod, Paris, 1958. MR Zbl

[Lin et al. 2011] Y. Lin, L. Lu, and S.-T. Yau, "Ricci curvature of graphs", *Tohoku Math. J.* (2) **63**:4 (2011), 605–627. MR Zbl

[Lin et al. 2014] Y. Lin, L. Lu, and S.-T. Yau, "Ricci-flat graphs with girth at least five", *Comm. Anal. Geom.* **22**:4 (2014), 671–687. MR Zbl

[Lott and Villani 2009] J. Lott and C. Villani, "Ricci curvature for metric-measure spaces via optimal transport", *Ann. of Math.* (2) **169**:3 (2009), 903–991. MR Zbl

[Münch and Wojciechowski 2019] F. Münch and R. K. Wojciechowski, "Ollivier Ricci curvature for general graph Laplacians: heat equation, Laplacian comparison, non-explosion and diameter bounds", *Adv. Math.* **356** (2019), 106759, 45. MR Zbl

[Ollivier 2009] Y. Ollivier, "Ricci curvature of Markov chains on metric spaces", *J. Funct. Anal.* **256**:3 (2009), 810–864. MR Zbl

[Ollivier 2010] Y. Ollivier, "A survey of Ricci curvature for metric spaces and Markov chains", pp. 343–381 in *Probabilistic approach to geometry*, edited by M. Kotani et al., Adv. Stud. Pure Math. **57**, Math. Soc. Japan, Tokyo, 2010. MR Zbl

[Smith 2014] J. D. H. Smith, "Ricci curvature, circulants, and a matching condition", *Discrete Math.* **329** (2014), 88–98. MR Zbl

[Sturm 2006] K.-T. Sturm, "On the geometry of metric measure spaces, I", *Acta Math.* **196**:1 (2006), 65–131. MR Zbl

Received: 2019-08-26 Revised: 2020-03-11 Accepted: 2020-08-06

vbonini@calpoly.edu *Mathematics Department, California Polytechnic State University, San Luis Obispo, CA, United States*

ccarro07@calpoly.edu *Mathematics Department, California Polytechnic State University, San Luis Obispo, CA, United States*

udinh@calpoly.edu *Mathematics Department, California Polytechnic State University, San Luis Obispo, CA, United States*

sydney.dye24@gmail.com *Mathematics Department, California Polytechnic State University, San Luis Obispo, CA, United States*

jmfreder@calpoly.edu *Mathematics Department, California Polytechnic State University, San Luis Obispo, CA, United States*

epearse@calpoly.edu *Mathematics Department, California Polytechnic State University, San Luis Obispo, CA, United States*

INVOLVE 13:4 (2020)
https://doi.org/10.2140/involve.2020.13.577

On equidistant polytopes in the Euclidean space

Csaba Vincze, Márk Oláh and Letícia Lengyel

(Communicated by Michael Dorff)

An equidistant polytope is a special equidistant set in the space $\mathbb{R}^n$ all of whose boundary points have equal distances from two finite systems of points. Since one of the finite systems of the given points is required to be in the interior of the convex hull of the other one, we can speak about inner and outer focal points of the equidistant polytope. It is of type (q, p), where q is the number of the outer focal points and p is the number of the inner focal points. The equidistancy is the generalization of convexity because a convex polytope can be given as an equidistant polytope of type $(q, 1)$, where $q \geq n + 1$. We present some general results about the basic properties of the equidistant polytopes: convex components, graph representations, connectedness, correspondence to the Voronoi decomposition of the space etc. In particular, we are interested in equidistant polytopes of dimension 2 (equidistant polygons). Equidistant polygons of type $(3, 2)$ will be characterized in terms of a constructive (ruler-and-compass) process to recognize them. In general they are pentagons with exactly two concave angles such that the vertices at which the concave angles appear are joined by an inner diagonal related to the adjacent sides of the polygon in a special way via the three reflections theorem for concurrent lines. The last section is devoted to some special arrangements of the focal points to get the concave quadrangles as equidistant polygons of type $(3, 2)$.

1. Introduction

Let $K \subset \mathbb{R}^n$ be a subset in the Euclidean coordinate space. The distance between a point $X \in \mathbb{R}^n$ and K is measured by the usual infimum formula

$$d(X, K) := \inf\{d(X, Y) \mid Y \in K\}.$$

The equidistant set of K and $L \subset \mathbb{R}^n$ is defined as

$$\{K = L\} := \{X \in \mathbb{R}^n \mid d(X, K) = d(X, L)\}.$$

MSC2010: 51M04.

Keywords: Euclidean geometry, convex geometry, equidistant sets.

Vincze Csaba is supported by the EFOP-3.6.2-16-2017-00015 project. The project was supported by the European Union, cofinanced by the European Social Fund. Lengyel Letícia is supported by the University of Debrecen (Summer Grant 2019).

Since the classical conics can also be given in this way [Ponce and Santibáñez 2014], the equidistant sets are their generalizations: K and L are called the focal sets. Moreover, any convex polytope can be given as an equidistant set with finitely many focal points in K and L, respectively [Vincze 2017]; see also [Vincze and Oláh 2019]. Therefore the equidistancy is the generalization of convexity. In a similar way we can speak about an equidistant function [Vincze et al. 2018a] by requiring its epigraph to be an equidistant body:

$$\{K \le L\} := \{X \in \mathbb{R}^n \mid d(X, K) \le d(X, L)\}.$$

In the case of singletons we are going to use the applicable shortcut notation

$$\{X \le L\} := \{\{X\} \le L\}, \quad \{X \le Y\} := \{\{X\} \le \{Y\}\}, \quad \text{etc.} \quad (X, Y \in \mathbb{R}^n).$$

The investigation of equidistant sets with finitely many focal points is motivated by a continuity theorem [Ponce and Santibáñez 2014]: if K and L are disjoint compact subsets in the space and $K_n \to K$ and $L_n \to L$ are convergent sequences of nonempty compact subsets with respect to the Hausdorff metric, then $\{K_n = L_n\}$ is a convergent sequence tending to $\{K = L\}$ with respect to the Hausdorff metric in any bounded region of the space. For the Hausdorff distance between compact sets in $\mathbb{R}^n$, see, e.g., [Lay 1982].

The points of an equidistant set are difficult to determine in general because there are no simple formulas to compute the distance between a point and a set. The continuity theorem allows us to simplify the general problem by using the approximation $\{K_n = L_n\} \approx \{K = L\}$, where $K_n \subset K$ and $L_n \subset L$ are finite subsets. In case of finite focal sets in two dimensions, the equidistant points can be characterized in terms of computable constants and parametrization [Vincze et al. 2018b] (the paper contains a MAPLE implementation as well as an alternative of the error estimation process for quasiequidistant points suggested by [Ponce and Santibáñez 2014] for the computer simulation). For some general investigations of equidistant sets in metric spaces we can refer to the fundamental works [Loveland 1976; Wilker 1975].

2. Convex components, graph representations and connectedness

Lemma 1. *Let K and L be nonempty compact subsets in the Euclidean coordinate space $\mathbb{R}^n$. The equidistant body*

$$\{K \le L\} := \{X \in \mathbb{R}^n \mid d(X, K) \le d(X, L)\}$$

can be expressed as the union

$$\{K \le L\} = \bigcup_{X \in K} \{X \le L\}.$$

In particular,

$$\{K_1 \cup K_2 \le L\} = \{K_1 \le L\} \cup \{K_2 \le L\}.$$

Proof. If $Z \in \{K \le L\}$ then $d(Z, K) \le d(Z, L)$ and, by the compactness (especially, the closedness) there exists a point $X \in K$, at which the minimal distance is attained:

$$d(Z, X) = d(Z, K) \le d(Z, L) \quad \Longrightarrow \quad Z \in \{X \le L\}.$$

Conversely, for any $X \in K$, we have $d(Z, K) \le d(Z, X)$; i.e., if $Z \in \{X \le L\}$, then

$$d(Z, K) \le d(Z, X) \le d(Z, L)$$

and $Z \in \{K \le L\}$, as was to be proved. □

The previous result motivates us to formulate some simple observations about the equidistant bodies of the form $\{X \le L\}$.

Lemma 2. *For any $X \in \mathbb{R}^n$ the set $\{X \le L\}$ is convex. If X is not an accumulation point of L then $\{X \le L\}$ is of dimension n.*

Proof. The convexity follows from the half-space intersection formula

$$\{X \le L\} = \bigcap_{Y \in L} \{X \le Y\}, \tag{1}$$

where $\{X \le Y\}$ is the closed half-space bounded by the perpendicular bisector of the segment XY containing X. On the other hand, if X is not an accumulation point of L, then it is an outer point or an isolated point. In the case of an outer point, $d(X, L) > 0$ and a continuity argument shows that X is an interior point of $\{X \le L\}$. Otherwise (in the case of an isolated point of L)

$$d(X, Z) = d(Z, L)$$

for any element Z in a sufficiently small open neighborhood of X. Therefore X is an interior point of $\{X \le L\}$. □

Remark 3. Note that the converse does not hold in general: if $X = (0, 0)$, $L = \{(0, 1/n) \mid n \in \mathbb{N}\}$ then

$$\{X \le L\} = \{(x, y) \mid x \le 0\}$$

is of dimension 2 but X is an accumulation point of L. It can be easily seen that:

- If X is an inner point of L then $\{X \le L\} = \{X\}$.
- If X is an outer point or an isolated point of L then X is an inner point of $\{X \le L\}$.
- If X is an accumulation point of L then $-X + \{X \le L\}$ is a subset in the regular normal cone [Rockafellar and Wets 1998] of the topological closure of L at the point X containing proximal normals of the form $Z - X$ ($Z \in \{X \le L\}$).

By Lemma 1 any equidistant body can be expressed as the union of convex subsets determined by its focal points. Following the steps in the proof we also have that

$$\{K < L\} := \{X \in \mathbb{R}^n \mid d(X, K) < d(X, L)\}$$

can be expressed as the union

$$\{K < L\} = \bigcup_{X \in K} \{X < L\}$$

and, consequently,

$$\begin{aligned}\{K \le L\} &= \bigcup_{X \in K} \{X \le L\} = \bigcup_{X \in K} \overline{\{L < X\}} \\ &= \overline{\bigcap_{X \in K} \{L < X\}} = \overline{\bigcap_{X \in K} \bigcup_{Y \in L} \{Y < X\}} = \bigcup_{X \in K} \bigcap_{Y \in L} \{X \le Y\}.\end{aligned}$$

On the other hand

$$\overline{\{K \le L\}} = \{L < K\} = \bigcup_{Y \in L} \{Y < K\} = \bigcup_{Y \in L} \overline{\{K \le Y\}} = \overline{\bigcap_{Y \in L} \{K \le Y\}}$$

$$\Longrightarrow \quad \{K \le L\} = \bigcap_{Y \in L} \{K \le Y\} = \bigcap_{Y \in L} \bigcup_{X \in K} \{X \le Y\}.$$

In particular,

$$\{K \le L_1 \cup L_2\} = \{K \le L_1\} \cap \{K \le L_2\}. \tag{2}$$

Lemma 4. *Let K and L be nonempty compact subsets in the Euclidean coordinate space $\mathbb{R}^n$. The equidistant body*

$$\{K \le L\} := \{X \in \mathbb{R}^n \mid d(X, K) \le d(X, L)\}$$

is bounded if and only if K is in the interior of the convex hull of L.

Proof. Suppose that the equidistant body is bounded and $X_1 \in K$ such that X_1 is not in the interior of the convex hull of L. Consider the closest point Y_1 of $\operatorname{conv} L$ to X_1. It is uniquely determined. Taking the supporting hyperplane H_1 to $\operatorname{conv} L$ at Y_1 such that

- it is perpendicular to the segment X_1Y_1 in case of $X_1 \neq Y_1$ (see Figure 1),
- it is an arbitrary supporting hyperplane in case of $X_1 = Y_1$,

the half-space containing the convex hull of L is called positive. Its complement is called negative. Choosing a point Q of the ray emanating from X_1 in the negative open half-space along the orthogonal direction to H_1, it can be easily seen that Y_1, X_1 and Q are collinear, i.e.,

$$d(Q, K) \le d(Q, X_1) \le d(Q, Y_1) = \inf_{Y \in \operatorname{conv} L} d(Q, Y) \le \inf_{Y \in L} d(Q, Y) = d(Q, L)$$

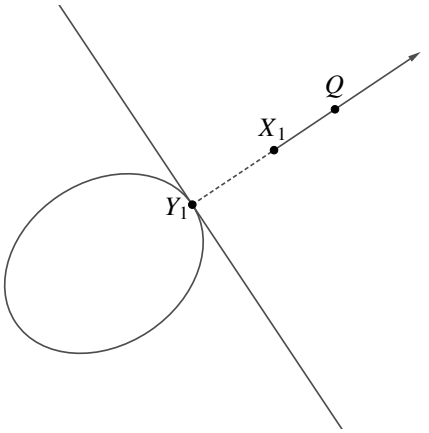


Figure 1. The proof of Lemma 4.

because of $L \subset \operatorname{conv} L$. Therefore $Q \in \{K \leq L\}$ but Q can tend to infinity. It is a contradiction. To prove the converse statement suppose that K is contained in the interior of the convex hull of L. By Lemma 1, for any $Q \in \{K \leq L\}$ we have $Q \in \{X \leq L\}$ for some point X in K. This means that

$$d(Q, X) \leq d(Q, Y) \quad (Y \in L).$$

Taking the square of both sides we have

$$\langle Q, Y - X\rangle \leq \tfrac{1}{2}|Y|^2 - \tfrac{1}{2}|X|^2 \leq c,$$

where the upper bound $c > 0$ can be chosen independently of Q due to the boundedness of K and L. Therefore Q/c is an element in the polar body $(-X + \operatorname{conv} L)^*$. The translated set $-X + \operatorname{conv} L$ contains the origin in its interior because $X \in K$ and K is in the interior of the convex hull of L. Taking a sufficiently small radius $r_X > 0$ we have

$$\begin{aligned} r_X D \subset -X + \operatorname{conv} L \quad &\Longrightarrow \quad Q/c \in (-X + \operatorname{conv} L)^* \subset (r_X D)^* = D/r_X \\ &\Longrightarrow \quad Q \in cD/r_X, \end{aligned}$$

where D is the closed unit ball around the origin in the space. Using a compactness argument, it follows that

$$r := \inf_{X \in K} \sup\{r_X \mid r_X D \subset -X + \operatorname{conv} L\} = \inf_{X \in K} \sup\{r_X \mid X + r_X D \subset \operatorname{conv} L\} > 0;$$

i.e., $Q \in cD/r$ for any $Q \in \{K \leq L\}$. □

2.1. ***The graph representation of equidistant bodies with finitely many focal points.*** In what follows some basic properties (connectedness) of an equidistant

body will be investigated by the pairwise comparison of the convex components in the union

$$\{K \leq L\} = \bigcup_{i=1}^{p} \{X_i \leq L\},$$

where $K = \{X_1, \ldots, X_p\}$, $L = \{Y_1, \ldots, Y_q\}$ are finite sets. It is motivated by the difficulties of the description of the geometric relationships among the focal points. The pairwise comparison of simple configurations seems to be a more effective way according to the algorithmic methods as well (see, e.g., the finite version of Helly's theorem).

Definition 5. The vertices of the graph representation G of the equidistant body $\{K \leq L\}$ are the elements of K and there is an edge between X_i and X_j $(i \neq j)$ if and only if

$$\{X_i \leq L\} \cap \{X_j \leq L\} \neq \varnothing.$$

The weight of the edge $X_i X_j$ is

$$w_{ij} = \dim\{X_i \leq L\} \cap \{X_j \leq L\}.$$

Since

$$\{X_i \leq L\} \overset{(2)}{=} \bigcap_{k=1}^{q} \{X_i \leq Y_k\},$$

the existence of the edge between X_i and X_j can be checked algorithmically by the finite version of Helly's theorem.

Theorem 6. *The equidistant body $\{K \leq L\}$ is connected if and only if its graph representation is connected.*

Proof. Suppose that $G = A \cup B$, where A and B are disjoint subsets of the vertices such that there is no edge with endpoints in A or B, respectively. Taking the sets

$$M = \bigcup_{X_i \in A} \{X_i \leq L\} \quad \text{and} \quad N = \bigcup_{X_i \in B} \{X_i \leq L\}$$

we have that M and N are closed disjoint subsets of the equidistant body and $\{K \leq L\} = M \cup N$ because $G = A \cup B$. If the equidistant body is connected then one of the sets, say M must be empty. Thus so is A; i.e., the graph representation is connected. Conversely, if G is connected then we have a sequence of edges from X_i to X_j for any pair of indices $i \neq j$: $X_i - X_{k_1} - \cdots - X_{k_m} - X_j$. The continuous path connecting $P \in \{X_i \leq L\}$ with $Q \in \{X_j \leq L\}$ can be constructed as follows: the first step is to join P with X_i (they are in the same convex component), the second step is to join X_i with a point

$$X_{ik_1} \in \{X_i \leq L\} \cap \{X_{k_1} \leq L\} \tag{3}$$

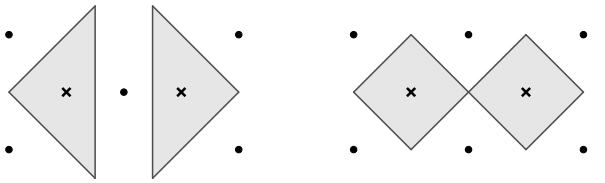


Figure 2. Some disconnected cases.

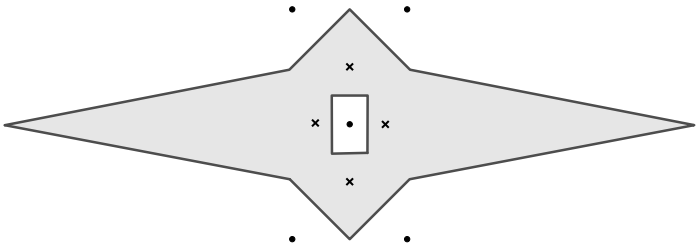


Figure 3. Disconnected components of the complement of an equidistant body.

and, finally, we can join X_{ik_1} with X_{k_1} because they are in the same convex component. The polygonal chain constructed by the successive application of these steps joins P with Q. Therefore the body is arcwise-connected; i.e., it is connected. □

According to the argument in the proof of Theorem 6, the connectedness and the arcwise-connectedness are equivalent for an equidistant body.

Corollary 7. *The equidistant body $\{K \leq L\}$ is arcwise-connected if and only if its graph representation is connected.*

Proof. If the equidistant body is arcwise-connected then it is connected. Thus so is its graph representation. Conversely, if G is connected then we can follow the process in the proof of the previous theorem to construct a polygonal chain between any pair of points P and Q in $\{K \leq L\}$. □

Corollary 8. *A disconnected equidistant body $\{K \leq L\}$ is the disjoint union of equidistant bodies of type $\{K_i \leq L\}$, where $K = K_1 \cup K_2 \cup \cdots$ is the disjoint union of subsets in K corresponding to the connected components of the graph representation.*

Some disconnected cases are illustrated in Figure 2: the equidistant body is disconnected (left); the equidistant body is connected but its interior is not (right). Figure 3 shows the case of a disconnected complement.

Definition 9. The graph representation G of the equidistant body $\{K \leq L\}$ is disconnected with respect to the weight w if $G = A \cup B$, where A and B are

nonempty disjoint subsets of the vertices such that there are no edges of weight greater than or equal to w with endpoints in A and B, respectively. Otherwise G is connected with respect to the weight w.

Theorem 10. *The interior of the equidistant body $\{K \leq L\}$ is connected if and only if its graph representation is connected with respect to the weight $n-1$.*

Proof. Suppose that the graph representation is connected with respect to the weight $n-1$ and let us modify the proof of Theorem 6 by changing the intermediate points $X_{ik_1}, X_{k_1k_2}, \ldots, X_{k_mj}$ such that they belong to the interior of the intersection of the corresponding convex components (an edge of maximal weight) or they are in the relative interior of the adjacent faces of dimension $n-1$. The modification gives a continuous path (polygonal chain) from X_i to X_j $(i \neq j)$ in the interior of the equidistant body. Since the arcwise-connectedness implies the connectedness, we are done. Conversely, suppose that the interior of the equidistant body $\{K \leq L\}$ is connected; i.e., it is arcwise-connected because the connectedness and the arcwise-connectedness are equivalent in case of open sets. To prove the connectedness of the graph representation with respect to the weight $n-1$ we are going to construct a sequence of edges of weight at least $n-1$ between X_i and X_j $(i \neq j)$. If $X_j \in \{X_i \leq L\}$ then we are done because $\dim\{X_i \leq L\} \cap \{X_j \leq L\} = n$. Indeed, since a finite set has no accumulation points, Lemma 2 implies that X_j is in the interior of $\{X_j \leq L\}$. In particular, each convex component is of dimension n. Finally, if a convex set of dimension n intersects the interior of another one then the intersection is of dimension n. Otherwise consider a continuous path $\widehat{X_iX_j}$ from X_i to X_j in the interior of the equidistant body and choose a common point Z of $\widehat{X_iX_j}$ with the boundary of $\{X_i \leq L\}$. Since Z is an interior point, we can suppose — without loss of generality — that Z is lying on the face F^i_{n-1} of dimension $n-1$ of the convex component $\{X_i \leq L\}$. Let $\varepsilon > 0$ be small enough and $B_Z(\varepsilon) \subset \operatorname{int}\{K \leq L\}$, where $B_Z(\varepsilon)$ is the open ball around Z with radius ε. Then $B_Z(\varepsilon) \cap F^i_{n-1}$ must be covered by the finite collection $\{\{X_k \leq L\} \mid k \neq i\}$. Condition $\dim B_Z(\varepsilon) \cap F^i_{n-1} = n-1$ implies that there must be at least one intersection $\{X_i \leq L\} \cap \{X_{k_1} \leq L\}$ of dimension at least $n-1$. Therefore w_{ik_1} (the weight of the edge $X_iX_{k_1}$) is at least $n-1$. Repeating the algorithm along the arc $\widehat{X_{k_1}X_j}$ we are done in finitely many steps. □

2.2. *Equidistant polytopes.* In what follows we are going to define the notion of equidistant polytopes. Some natural requirements are the connectedness (see Corollary 8), the boundedness (see Lemma 4) and the finiteness of the focal sets.

Definition 11. Let K and $L \subset \mathbb{R}^n$ be nonempty disjoint, finite sets and suppose that K is a subset in the interior of the convex hull of L. The equidistant body $\{K \leq L\}$ is called an equidistant polytope if both its interior and its complement are connected. The equidistant polytope is of type (q, p), where $|L| = q$ and $|K| = p$.

Corollary 12. *The boundary $\{K = L\}$ of an equidistant polytope is connected.*

The equidistant polytopes of type $(q, 1)$ are convex polytopes. It is a direct consequence of the half-space intersection formula (1) and Lemma 4: an equidistant polytope of type $(q, 1)$ is a nonempty, compact intersection of finitely many closed half-spaces. Since any convex polytope can be given as an equidistant polytope of type $(q, 1)$, the equidistancy is the generalization of the convexity; see [Vincze 2017; Vincze and Oláh 2019]. In a similar way we can speak about equidistant functions [Vincze et al. 2018a] by requiring their epigraphs to be equidistant bodies. Let $\{K \leq L\}$ be an equidistant polytope; since K must be in the interior of the convex hull of L, the set of the outer focal points must contain at least $n + 1$ points.

Lemma 13. *If K and $L \subset \mathbb{R}^n$ are nonempty disjoint finite sets, $|L| = n + 1$ and K is a subset in the interior of the convex hull of L, then the equidistant body $\{K \leq L\}$ is a star-shaped set.*

Proof. The idea is to provide the existence of a closed ball strictly separating K and L in the sense that the focal points of K are inside but the focal points of L are outside. [1] Then the center X_* of the ball satisfies the inequality

$$\max\{d(X_*, X_1), \ldots, d(X_*, X_p)\} < d(X_*, L) = \min\{d(X_*, Y_1), \ldots, d(X_*, Y_q)\},$$

where $X_1, \ldots, X_p$ are the points in K and $Y_1, \ldots, Y_q$ are the points in L. Therefore, by a continuity argument, X_* is an interior point of every convex component $\{X_i \leq L\}$, where $i = 1, \ldots, p$. By Lemma 1, this means that the equidistant body $\{K \leq L\}$ is star-shaped with respect to any point in an open ball around X_* with a sufficiently small radius. If $q = n+1$ then L is a simplex and we can easily construct a strictly separating ball by a slight decreasing of the radius of its circumscribed sphere. □

Corollary 14. *If K and $L \subset \mathbb{R}^n$ are nonempty disjoint finite sets, $|L| = n+1$ and K is a subset in the interior of the convex hull of L, then the equidistant body $\{K \leq L\}$ is an equidistant polytope.*

2.3. ***Voronoi decomposition and its correspondence to the equidistant bodies.*** One of the most important applications of equidistant bodies of the form $\{X \leq L\}$ is the Voronoi decomposition of the space. Let $K := \{X_1, \ldots, X_m\} \subset \mathbb{R}^n$ be a finite set containing different elements and consider the equidistant bodies

$$V(X_i, K) := \{X_i \leq K \setminus \{X_i\}\} \quad (i = 1, \ldots, m).$$

[1] The combinatorial criteria of the existence of such a separating ball can be formulated as a Kirchberger-type theorem [Lay 1982]: if for any subset $T \subset K \cup L$ containing $n + 3$ points there is a ball strictly separating $T \cap K$ and $T \cap L$ then there is a ball strictly separating K and L.

It is clear that they are n-dimensional convex subsets (Voronoi cells) such that

$$\mathbb{R}^n = \bigcup_{i=1}^{m} V(X_i, K)$$

and

$$\operatorname{int} V(X_i, K) \cap \operatorname{int} V(X_j, K) = \varnothing \quad (i \neq j = 1, \ldots, n).$$

The set $V(X_i, K)$ contains the points in the space, where the distance $d(X, K)$ is attained at X_i; i.e., X_i is the closest point of K to any $X \in V(X_i, K)$. The collection of the cells $V(X_1, K), \ldots, V(X_m, K)$ is called the Voronoi decomposition of the space with respect to the set K.

Corollary 15. *If K and L are nonempty finite, disjoint subsets containing different elements then*

$$\{K \leq L\} = \bigcup_{X_i \in K} V(X_i, K \cup L).$$

3. Equidistant polygons in the plane: the hypergraph representation and the maximal number of the vertices

The following investigations are motivated by the discrete version of the problem posed in [Ponce and Santibáñez 2014]: *characterize all closed sets of the plane that can be realized as the equidistant set of two connected disjoint closed sets.*

3.1. ***The hypergraph representation of equidistant polygons.*** Let K and $L \subset \mathbb{R}^2$ be nonempty disjoint finite subsets in the plane such that K is contained in the interior of the convex hull of L. In what follows we are going to estimate the maximal number of the edges (vertices) of an equidistant polygon in terms of the number of elements in the focal sets. Using a continuity argument it can be easily seen that decreasing the number of vertices is impossible by a slight modification of the position[2] of the focal points (increasing the number of vertices is possible). This means that the regularity conditions

(C1) there are no collinear triplets among the points of $K \cup L$,

(C2) there are no concircular quadruples among the points of $K \cup L$

can be supposed without loss of generality. The regularity conditions do not imply in general that we have an equidistant polygon as Figure 3 shows by a slight modification of the focal sets. In the special case of two dimensions, the connectedness of the interior of an equidistant polygon provides that there are

[2]Since the convex components are nonempty compact intersections of finitely many half-spaces, they depend continuously on the position of the focal points in any bounded region of the space. So does their finite union. Therefore vertices (breakages along the boundary) cannot be straightened by a slight modification of the position of the focal points but straight line segments can be broken.

no self-intersections of its boundary and the connectedness of its complement provides that there are no holes in its interior. Using the boundedness criterion ($K \subset \operatorname{int}\operatorname{conv} L$) and the finiteness of the focal sets, Corollary 12 implies that the boundary of an equidistant polygon in the plane is a simple closed polygonal chain (the edges belong to the perpendicular bisectors of the elements in the focal sets K and L, respectively). Therefore an equidistant polygon in the plane is a Jordan polygon (see the Jordan curve theorem).

Definition 16. The 3-uniform hypergraph representation of an equidistant polygon satisfying (C1) and (C2) consists of the vertices $K \cup L$ and the edges are the triplets of the focal points provided that the circle determined by them does not contain any focal point in its interior. Edges of types $\{X_{i_1}, X_{i_2}, X_{i_3}\}$ and $\{Y_{j_1}, Y_{j_2}, Y_{j_3}\}$ are called monochromatic. The colored edges are of types $\{X_{i_1}, Y_{j_1}, X_{i_2}\}$ and $\{Y_{j_1}, X_{i_1}, Y_{j_2}\}$. The weight of a colored edge is the angle $\delta_{i_1 i_2}^{j_1} = \angle X_{i_1} Y_{j_1} X_{i_2}$ or $\omega_{j_1 j_2}^{i_1} = \angle Y_{j_1} X_{i_1} Y_{j_2}$.

To prevent the inclusion of focal points in the interior of the circle determined by a colored edge it is sufficient and necessary for the weight to satisfy

$$\begin{aligned}\delta_{i_1 i_2}^{j_1} &= \max\{\angle X_{i_1} Z X_{i_2} \mid Z \in K \cup L \text{ and } Z \in H_K^+(i_1, j_1, i_2)\} \\ &\leq \pi - \max\{\angle X_{i_1} Z X_{i_2} \mid Z \in K \cup L \text{ and } Z \in H_K^-(i_1, j_1, i_2)\},\end{aligned}$$

where the open half-plane $H_K^+(i_1, j_1, i_2)$ is bounded by the line $X_{i_1} X_{i_2}$ and it contains the point Y_{j_1}, $H_K^-(i_1, j_1, i_2)$ is the opposite open half-plane, or

$$\begin{aligned}\omega_{j_1 j_2}^{i_1} &= \max\{\angle Y_{j_1} Z Y_{j_2} \mid Z \in K \cup L \text{ and } Z \in H_L^+(j_1, i_1, j_2)\} \\ &\leq \pi - \max\{\angle Y_{j_1} Z Y_{j_2} \mid Z \in K \cup L \text{ and } Z \in H_L^-(j_1, i_1, j_2)\},\end{aligned}$$

where the open half-plane $H_L^+(j_1, i_1, j_2)$ is bounded by the line $Y_{j_1} Y_{j_2}$ and it contains the point X_{i_1}; $H_L^-(j_1, i_1, j_2)$ is the opposite open half-plane.

Corollary 17. *Consider the hypergraph representation of an equidistant polygon satisfying* (C1) *and* (C2). *The centers of the circles determined by monochromatic edges of types* $\{X_{i_1}, X_{i_2}, X_{i_3}\}$ *and* $\{Y_{j_1}, Y_{j_2}, Y_{j_3}\}$ *are interior and exterior points of the equidistant polygon, respectively. The centers of the circles determined by colored edges are the vertices of the equidistant polygon.*

Proof. If $\{X_{i_1}, X_{i_2}, X_{i_3}\}$ is a monochromatic edge then there are no focal points in the interior of the circle determined by X_{i_1}, X_{i_2} and X_{i_3}. The existence of such a circle is due to (C1). Let $r_{i_1 i_2 i_3}$ and $X_{i_1 i_2 i_3}$ be the radius and the center of the circle, respectively. We have

$$d(X_{i_1 i_2 i_3}, K) = r_{i_1 i_2 i_3} < \min\{d(X_{i_1 i_2 i_3}, Y_1), \ldots, d(X_{i_1 i_2 i_3}, Y_q)\} = d(X_{i_1 i_2 i_3}, L),$$

where the strict inequality is due to (C2). Therefore $X_{i_1 i_2 i_3}$ is in the interior of $\{K \leq L\}$. The argument is similar in the case of a monochromatic edge $\{Y_{j_1}, Y_{j_2}, Y_{j_3}\}$.

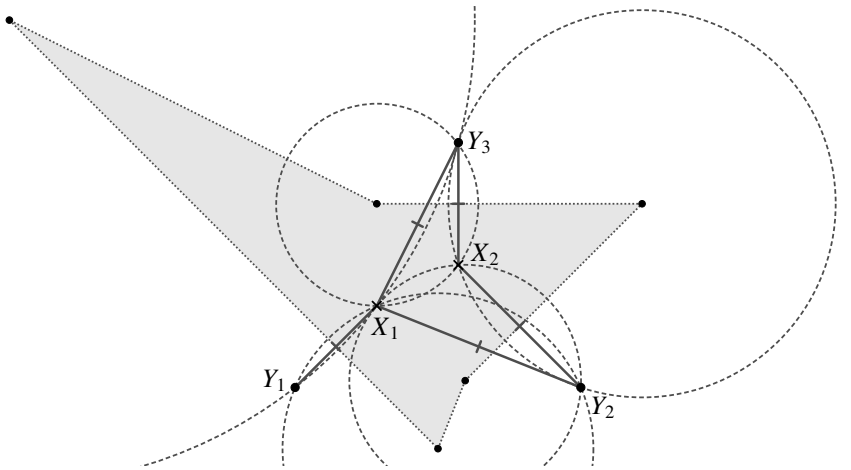


Figure 4. The bigraph of the pentagon: the proof of Lemma 18.

Taking a colored edge $\{X_{i_1}, Y_{j_1}, X_{i_2}\}$, let $r_{i_1 j_1 i_2}$ and $V_{i_1 j_1 i_2}$ be the radius and the center of the circle determined by the points of the triplet, respectively. The existence of such a circle is due to (C1). We have

$$d(V_{i_1 j_1 i_2}, K) = r_{i_1 j_1 i_2} = d(V_{i_1 j_1 i_2}, Y_{j_1}) = d(V_{i_1 j_1 i_2}, L).$$

This means that $V_{i_1 j_1 i_2}$ is an equidistant point of K and L. The perpendicular bisectors of the chords $X_{i_1} Y_{j_1}$ and $X_{i_2} Y_{j_1}$ intersect each other at $V_{i_1 j_1 i_2}$. According to (C2) there are no equidistant points in a sufficiently small open neighborhood of $V_{i_1 j_1 i_2}$ except the points of the perpendicular bisectors. They determine a concave angle because X_{i_1} and X_{i_2} are automatically in the interior of $\{K \leq L\}$. The argument is similar in case of a colored edge $\{Y_{j_1}, X_{i_1}, Y_{j_2}\}$. □

The colored edge $\{X_{i_1}, Y_{j_1}, X_{i_2}\}$ represents a single inner change in the sense that the vertex (the center of the circle determined by the elements of the triplet) is due to the change of the inner focal points (the outer focal point is the same). The circle is passing through exactly two of the inner and exactly one of the outer focal points (concave angles). The colored edge $\{Y_{j_1}, X_{i_1}, Y_{j_2}\}$ represents a single outer change in the sense that the vertex (the center of the circle determined by the elements of the triplet) is due to the change of the outer focal points (the inner focal point is the same). The circle is passing through exactly two of the outer and exactly one of the inner focal points (convex angles). Condition (C2) does not allow "double changes" in the sense that the vertex is due to the simultaneous change of the outer and the inner focal points. Such a kind of change will appear among the cases of the special arrangements of the focal points: Figure 8 shows a double change at V_1, single outer changes at V_2 and V_4, a single inner change at V_3.

Lemma 18. *An equidistant polygon of type* (q, p) *has at most* $p+q$ *vertices.*

Proof. Using a continuity argument it follows that decreasing the number of vertices is impossible by a slight modification of the position of the focal points (increasing the number of vertices is possible; see footnote 2). Therefore we can suppose that (C1) and (C2) are satisfied. Taking the hypergraph representation suppose that it is minimal in the sense that only the focal points belonging to colored edges are considered. This means that we have a finite chain of circles such that the centers form the vertices of the equidistant polygon in a given direction. The adjacent vertices correspond to adjacent circles having a common chord with endpoints $X_{i_m} \in K$ and $Y_{j_m} \in L$. Let us choose a starting vertex/circle. The following algorithm generates a bigraph (see Figure 4) with edges

(i) $e_1 := X_{i_1} Y_{j_1}$,

(ii) if $e_m := X_{i_m} Y_{j_m}$ and (X_{i_m}, Y_{j_m}, Z) determines the adjacent circle with respect to the given direction then

$$e_{m+1} = \begin{cases} X_{i_m} Z & \text{if } Z \in L, \\ Y_{j_m} Z & \text{if } Z \in K. \end{cases}$$

There is a one-to-one correspondence between the edges and the circles. Therefore the number of the edges equals to the number of the circles (the number of the vertices of the equidistant polygon). On the other hand, exactly one new element in $K \cup L$ appears in each step. This means that the number of the circles (the number of the vertices of the equidistant polygon) is less than or equal to $p+q$ as was to be proved. □

Remark 19. We can improve the estimation in the case of $p = 1$ as follows: the number of vertices satisfies

$$|(q, p)| = \begin{cases} q & \text{if } p = 1, \\ p+q & \text{otherwise} \end{cases}$$

provided that the hypergraph representation is minimal. Indeed, each point in the minimal representation $K \cup L$ appears in the matching process (i) and (ii) and each pair in the matching corresponds to a consecutive circle. The edges $e_1, \ldots, e_m, \ldots$ of the bigraph are orthogonal segments to the edges of the equidistant polygon.

4. Equidistant polygons of type (3, 2) in the plane: the generic case

Suppose that we have an equidistant polygon of type (3, 2) satisfying (C1) and (C2), $K = \{X_1, X_2\}$ and $L = \{Y_1, Y_2, Y_3\}$. Using Lemma 18 the maximal number of the vertices is 5 and it can be attained as we shall see. Let ω_{12}^i, ω_{23}^i and ω_{31}^i be the viewing angles under which the segments Y_1Y_2, Y_2Y_3 and Y_3Y_1 are visible from the

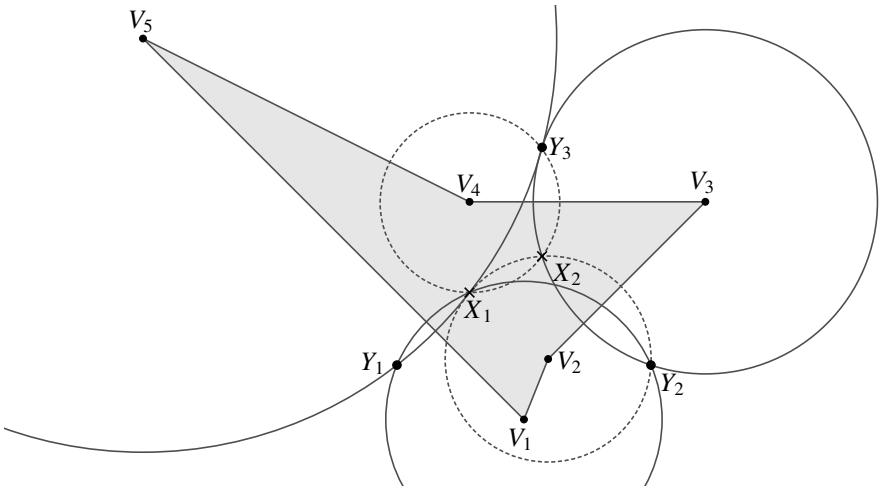


Figure 5. An equidistant polygon of type (3, 2): the generic case.

inner focal point X_i $(i = 1, 2)$. It is clear that

$$\omega_{31}^i = 360^\circ - \omega_{12}^i - \omega_{23}^i \quad (i = 1, 2). \tag{4}$$

We also introduce the viewing angle $\delta^j := \delta_{12}^j$ under which the segment X_1X_2 is visible from Y_j $(j = 1, 2, 3)$. According to condition (C2) we can suppose that

$$\omega_{12}^1 > \omega_{12}^2, \quad \omega_{23}^1 < \omega_{23}^2, \quad \omega_{31}^1 > \omega_{31}^2 \tag{5}$$

because $\omega_{23}^1 > \omega_{23}^2$ implies that $\omega_{31}^1 < \omega_{31}^2$ and the ordering (5) follows by changing the roles of Y_1 and Y_2. This means that $\{Y_1, X_1, Y_2\}$, $\{Y_2, X_2, Y_3\}$ and $\{Y_3, X_1, Y_1\}$ are colored edges in the hypergraph representation. They correspond to the convex angles of the equidistant polygon at V_1, V_3 and V_5 (Figure 5). What about the viewing angles δ^1, δ^2 and δ^3? Since condition (C1) is satisfied, the line X_1X_2 strictly separates two focal points from the third one, say Y_3. Using condition (C2) we can suppose that $\delta^1 < \delta^2$ and, consequently, the centers of the circles (colored edges) $\{X_1, Y_2, X_2\}$ and $\{X_1, Y_3, X_2\}$ are vertices of the equidistant polygon at which concave angles appear. Therefore we have a pentagon with exactly two concave angles at V_2 and V_4 (Figure 5). The focal points X_1 and X_2 are obviously symmetric about the line V_2V_4.

Lemma 20. *Consider a simple pentagon P with exactly two concave angles and let the vertices be labeled by A, B, C, D and E in the counterclockwise direction such that the concave angles are at B and D. If the auxiliary lines f_B and f_D are defined by*

$$\rho_{f_B} = \rho_{BA} \circ \rho_{BC} \circ \rho_{BD}, \quad \rho_{f_D} = \rho_{DE} \circ \rho_{DC} \circ \rho_{DB},$$

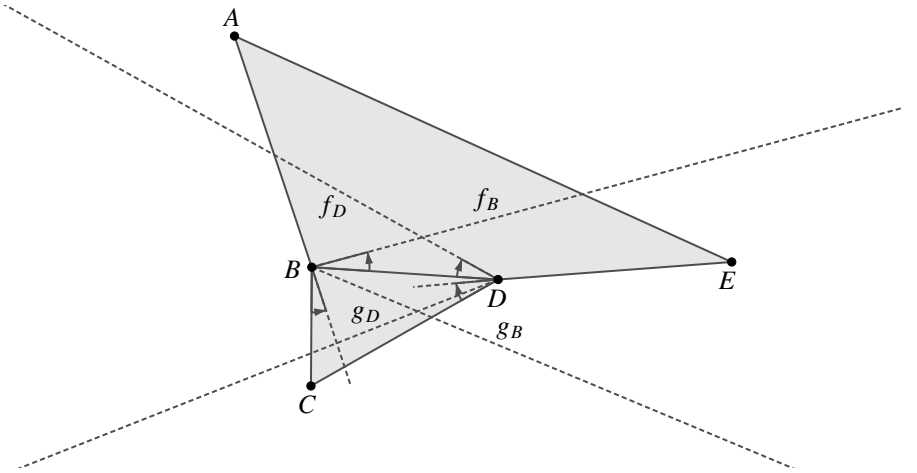


Figure 6. The proof of Lemma 20: pseudo inner focal points.

where ρ_{BA}, ρ_{BC}, ... denote the reflections about the lines determined by the indices, then f_B and f_D intersect each other on the side of the inner diagonal BD containing A.

Proof. First of all note that f_B and f_D are well-defined due to the three reflections theorem for concurrent lines. The theorem states that the composition of reflections about three concurrent lines is a reflection about a line passing through the common point. Figure 6 shows that the angle between the lines f_B and BD on the side of the inner diagonal BD containing A is just $\angle B - \pi$, where $\angle B$ is the concave angle of the polygon at B. In a similar way, $\angle D - \pi$ is the angle enclosed by f_D and DB on the side of DB containing A. Since

$$\angle A + \angle B + \angle C + \angle D + \angle E = 3\pi,$$

it follows that

$$(\angle B - \pi) + (\angle D - \pi) = \pi - (\angle A + \angle C + \angle E) < \pi$$

and the intersection point of f_B and f_D exists on the side of the inner diagonal BD containing A. □

Corollary 21. *Consider a simple pentagon P with exactly two concave angles and let the vertices be labeled by A, B, C, D and E in the counterclockwise direction such that the concave angles are at B and D. If the auxiliary lines g_B and g_D are defined by*

$$\rho_{g_B} = \rho_{BC} \circ \rho_{BA} \circ \rho_{BD}, \quad \rho_{g_D} = \rho_{DC} \circ \rho_{DE} \circ \rho_{DB},$$

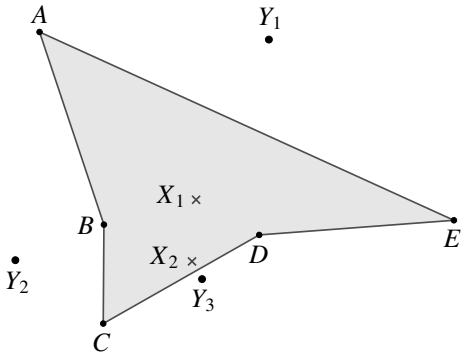


Figure 7. The proof of Theorem 3.

where ρ_{BC}, ρ_{BA}, ... denote the reflections about the lines determined by the indices, then g_B and g_E intersect each other on the side of the inner diagonal BD containing C.

Proof. Note that f_B and g_B (or f_D and g_D) are symmetric about the line BD because (for example) for any $G \in g_B$

$$\begin{aligned}\rho_{BD}(G) &= \rho_{BD} \circ \rho_{g_B}(G) = \rho_{BD} \circ \rho_{BC} \circ \rho_{BA} \circ \rho_{BD}(G)\\ &= \rho_{f_B}^{-1} \circ \rho_{BD}(G) = \rho_{f_B} \circ \rho_{BD}(G);\end{aligned}$$

i.e., $\rho_{BD}(G) \in f_B$ and vice versa. □

Definition 22. Let P be a simple pentagon with exactly two concave angles such that the vertices are labeled by A, B, C, D and E in the counterclockwise direction and the concave angles are at B and D. The intersection points $f_B \cap f_D$ and $g_B \cap g_D$ are called the pseudo inner focal points of P.

Theorem 23. *A simple pentagon is an equidistant polygon of type* (3, 2) *if and only if it has exactly two concave angles such that the vertices at which the concave angles appear are joined by an inner diagonal of the polygon and the pseudo inner focal points are in its interior.*

Proof. Suppose that P is a simple pentagon satisfying the conditions of the statement. Using the notation in Figure 6

$$X_1 := f_B \cap f_D, \quad X_2 = g_B \cap g_D;$$

they are symmetric about the line BD (see the proof of Corollary 21). The outer focal points are

$$Y_1 = \rho_{AE}(X_1), \quad Y_2 = \rho_{AB}(X_1) = \rho_{BC}(X_2), \quad Y_3 = \rho_{CD}(X_2) = \rho_{DE}(X_1);$$

see Figure 7. □

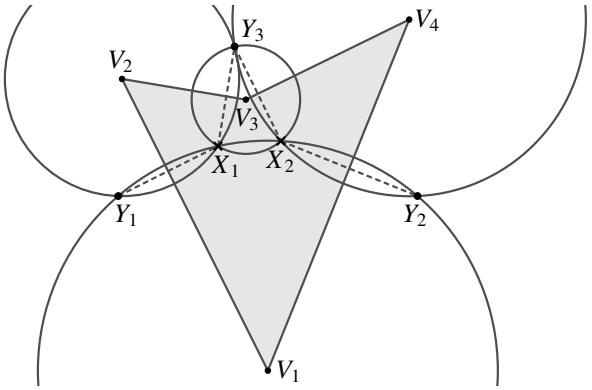


Figure 8. The case of concircular points.

5. Equidistant polygons of type (3, 2) in the plane: special arrangements of the focal points

5.1. ***The case of concircular points.*** In this case we have points in $K \cup L$, say X_1, Y_1, X_2 and Y_2 lying on the same circle. In particular, $\omega_{12}^1 = \omega_{12}^2$. Since the interior of the convex hull of L contains the points in K, all the focal points cannot be on the same circle. Without loss of generality we can suppose (by renumbering the inner focal points if necessary) that $\omega_{23}^1 < \omega_{23}^2$, as Figure 8 shows. Therefore $\omega_{31}^1 > \omega_{31}^2$ and there are convex angles at V_2 and V_4. Another convex angle is at the vertex V_1 due to the simultaneous change of the outer and the inner focal points. Since X_1, Y_1, X_2 and Y_2 are lying on the same circle, the viewing angles δ_1 and δ_2 are equal to each other and the secant line X_1X_2 strictly separates Y_1 and Y_2 from Y_3. This means that the center V_3 of the circle passing through the points X_1, X_2 and Y_3 is a vertex of the equidistant polygon at which a concave angle appears.

Theorem 24. *A simple concave quadrangle is an equidistant polygon of type* (3, 2).

Proof. Let the vertices of a simple concave quadrangle in the plane be labeled by A, B, C and D in the counterclockwise direction and suppose that the concave angle is at the vertex C. Let us introduce the auxiliary line f passing through the vertex C such that

$$\rho_f = \rho_{CD} \circ \rho_{CB} \circ \rho_{CA},$$

where ρ_{CD}, ρ_{CB}, ... denote the reflections about the lines determined by the indices. It is well-defined due to the three reflections theorem for concurrent lines. Since f passes through the vertex of the concave angle, it must contain points such that they are in the interior of the polygon together with their reflected pairs about the inner diagonal line (see Figure 9). Taking such a point X_1 we define

$$X_2 := \rho_{CA}(X_1).$$

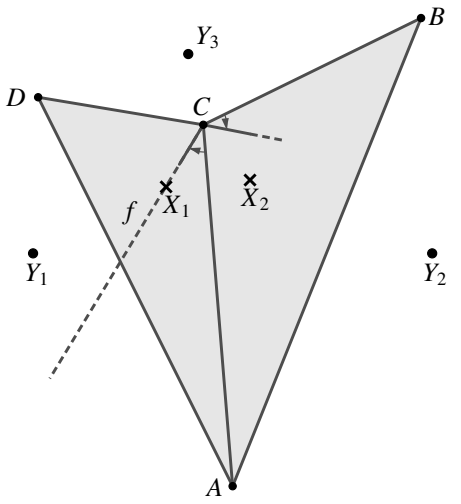


Figure 9. A simple concave quadrangle as an equidistant polygon of type (3, 2): Theorem 24.

According to the construction, $\rho_{CD}(X_1) = \rho_{CB}(X_2) = Y_3$. Finally we complete the set of the outer focal points by $\rho_{AD}(X_1) = Y_1$ and $\rho_{AB}(X_2) = Y_2$. □

In the proof of the previous theorem we can also consider the auxiliary line g determined by

$$\rho_g = \rho_{CB} \circ \rho_{CD} \circ \rho_{CA}$$

instead of f. The lines g and f are symmetric about the inner diagonal line AC because for any $G \in g$

$$\begin{aligned}\rho_{CA}(G) &= \rho_{CA} \circ \rho_g(G) = \rho_{CA} \circ \rho_{CB} \circ \rho_{CD} \circ \rho_{CA}(G)\\ &= \rho_f^{-1} \circ \rho_{CA}(G) = \rho_f \circ \rho_{CA}(G);\end{aligned}$$

i.e., $\rho_{CA}(G) \in f$ and vice versa. Since the inner focal points must be chosen symmetrically about the inner diagonal line, the role of these lines is also symmetric in the argumentation. Indeed, X_1X_2 is a common chord of the circles around V_1 and V_3 (Figure 8).

Corollary 25. *Any simple quadrangle is an equidistant polygon.*

Proof. Recall that convex quadrangles are equidistant polygons of type (4, 1); see [Vincze 2017]. Otherwise we can refer to Theorem 24. □

5.2. ***The case of collinear points.*** Suppose that one of the outer focal points, say Y_1, is collinear with X_1 and X_2. Since the inner focal points must be in the interior of the convex hull of the outer focal points, Y_2 and Y_3 must be strictly separated

by the line X_1X_2 and, consequently, the centers of the circles $\{X_1, X_2, Y_2\}$ and $\{X_1, X_2, Y_3\}$ are vertices of the equidistant polygon at which concave angles appear. It is the same situation as in the generic case.

5.3. ***Summary.*** We have proved that an equidistant polygon of type $(3, 2)$ in the plane belongs to one of the following classes:

- Simple concave quadrangles (four concircular focal points, the focal sets form one-parameter families as the point X_1 is moving along the auxiliary line f).
- Simple pentagons with exactly two concave angles such that the vertices at which the concave angles appear are joined by an inner diagonal and the pseudo inner focal points are in the interior of the pentagon. The pseudo inner focal points are constructed by the intersections of the lines substituting the adjacent sides and the inner diagonal at the vertices at which the concave angles appear via the three reflections theorem for concurrent lines (the focal sets are uniquely determined).

References

[Lay 1982] S. R. Lay, *Convex sets and their applications*, John Wiley & Sons, New York, 1982. MR Zbl

[Loveland 1976] L. D. Loveland, "When midsets are manifolds", *Proc. Amer. Math. Soc.* **61**:2 (1976), 353–360. MR Zbl

[Ponce and Santibáñez 2014] M. Ponce and P. Santibáñez, "On equidistant sets and generalized conics: the old and the new", *Amer. Math. Monthly* **121**:1 (2014), 18–32. MR Zbl

[Rockafellar and Wets 1998] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*, Grundlehren der Mathematischen Wissenschaften **317**, Springer, 1998. MR Zbl

[Vincze 2017] C. Vincze, "On convex closed planar curves as equidistant sets", preprint, 2017. arXiv

[Vincze and Oláh 2019] C. Vincze and M. Oláh, "Convex polytopes as equidistant sets in the space", preprint, 2019.

[Vincze et al. 2018a] C. Vincze, A. Varga, M. Oláh, and L. Fórián, "On computable classes of equidistant sets: equidistant functions", *Miskolc Math. Notes* **19**:1 (2018), 677–689. MR Zbl

[Vincze et al. 2018b] C. Vincze, A. Varga, M. Oláh, L. Fórián, and S. Lőrinc, "On computable classes of equidistant sets: finite focal sets", *Involve* **11**:2 (2018), 271–282. MR Zbl

[Wilker 1975] J. B. Wilker, "Equidistant sets and their connectivity properties", *Proc. Amer. Math. Soc.* **47** (1975), 446–452. MR Zbl

Received: 2019-09-02 Revised: 2020-07-20 Accepted: 2020-07-21

csvincze@science.unideb.hu *Institute of Mathematics, University of Debrecen, Debrecen, Hungary*

olma4000@gmail.com *Institute of Mathematics, Doctoral School of Mathematical and Computational Sciences, University of Debrecen, Debrecen, Hungary*

letikee123@gmail.com *Mathematics MSc, Institute of Mathematics, University of Debrecen, Debrecen, Hungary*

INVOLVE 13:4 (2020)
https://doi.org/10.2140/involve.2020.13.597

Polynomial values in Fibonacci sequences

Adi Ostrov, Danny Neftin, Avi Berman and Reyad A. Elrazik

(Communicated by Kenneth S. Berenhaut)

The only perfect powers in the Fibonacci sequence are 0, 1, 8, and 144, and in the Lucas sequence, the only perfect powers are 1 and 4. We prove that in sequences that follow the same recurrence relation of the Lucas and Fibonacci sequences, there are always only finitely many polynomial values $g(\mathbb{Z})$ for any polynomial g which is not equivalent to a Dickson polynomial.

1. Introduction

The Fibonacci $F^{0,1}$ and Lucas $F^{2,1}$ sequences are special cases of generalized Fibonacci sequences $F^{a,b} = \{F_n\}_{n=0}^{\infty},\ a, b \in \mathbb{Z}$, defined by the recurrence relation $F_{n+1} = F_n + F_{n-1}$, with initial values $F_0 = a,\ F_1 = b$.

Classical results show that 4 is the largest perfect square in the Lucas sequence and that 144 is the largest perfect square in the Fibonacci sequence [Cohn 1964]. Subsequently, much work was devoted to studying perfect powers in the Fibonacci and Lucas sequences, leading to their determination in [Bugeaud et al. 2006, §2]. This motivates the following natural question:

Question 1. For which polynomials $g \in \mathbb{Q}[x]$ does $F^{a,b},\ a, b \in \mathbb{Z}$, contain only finitely many values from $g(\mathbb{Z})$?

The following theorem answers the question when g is not a Dickson polynomial composed with linear polynomials. For $m \in \mathbb{Q}^{\times}$, denote by $D_{d,m} \in \mathbb{Q}[x]$ the Dickson polynomial of degree d, that is, the unique polynomial of degree d satisfying

$$D_{d,m}\left(x + \frac{m}{x}\right) = x^d + \frac{m^d}{x^d}.$$

Define the (normalized) Chebychev polynomial of degree d to be $T_d := D_{d,1}$.

MSC2010: 11B39, 11C08.
Keywords: recurrence sequence, polynomial values, Fibonacci.

Let φ denote the golden ratio $(1+\sqrt{5})/2$, and let $\chi_{a,b}$ represent the norm of $a+b\varphi \in \mathbb{Z}[\varphi]$, so that $\chi_{a,b} = N_{\mathbb{Z}[\varphi]/\mathbb{Z}}(a+b\varphi) = a^2+ab-b^2$.

Theorem 1. *Let $F^{a,b}$, $a, b \in \mathbb{Z}$, be a generalized Fibonacci sequence, and $g(x) \in \mathbb{Q}[x]$ a polynomial of degree $d > 1$. Assume that $g(x) \neq \alpha_{\chi,d,m} D_{d,m}(\mu(x))$ for all linear $\mu \in \mathbb{Q}[x]$ and $m \in \mathbb{Z} \setminus \{0\}$, where $\alpha_{\chi,d,m} = \pm\sqrt{-\chi/(5m^d)}$ for $\chi = \chi_{a,b}$ or $-\chi_{a,b}$. Then $F^{a,b}$ contains only finitely many values from $g(\mathbb{Z})$; that is, $\#(g(\mathbb{Z}) \cap F^{a,b}) < \infty$. In particular, $F^{a,b}$ contains only finitely many perfect squares.*

If $g(x)$ is of the form $\alpha_{\chi,d,m} D_{d,m}(\mu(x))$, then $\alpha_{\chi,d,m} \in \mathbb{Q}$, and hence -5χ is a square if d is even, and $-5m\chi$ is a square if d is odd. Also, flipping the sign of $\chi = \chi_{a,b}$ corresponds to shifting the sequence by one element, that is, $\chi_{b,a+b} = -\chi_{a,b}$.

We note that the sequence $F^{a,b}$ may indeed have infinite intersection with $\pm\alpha_{\chi,d,1} T_d(\mathbb{Z})$. For example, the Lucas sequence, which has $|\chi_{2,1}| = 5$ and satisfies the identity $L_{2n} = L_n^2 - (-1)^n \cdot 2$ [Cohn 1964, (3)], has infinite intersection with $g(\mathbb{Z})$ for $g(x) = T_2(x) = x^2 - 2$ and for $g(x) = -T_2(ix) = x^2 + 2$, where $i = \sqrt{-1}$. The Fibonacci sequence, which has $|\chi_{0,1}| = 1$ and satisfies the identity $F_{3n} = 5F_n^3 + 3(-1)^n F_n$ [Nagy et al. 2019, Table 2], has infinite intersection with $g(\mathbb{Z})$ for

$$g(x) = \frac{1}{\sqrt{5}} T_3(\sqrt{5}x) = 5x^3 - 3x \quad \text{and} \quad g(x) = -\frac{1}{\sqrt{-5}} T_3(\sqrt{-5}x) = 5x^3 + 3x.$$

Also note that in all examples the values $\{n : F_n \in g(\mathbb{Z})\}$ constitute the union of arithmetic progressions up to a finite set, as shown in the final assertion of [Corvaja and Zannier 1998, Theorem 2] (applied to $F(X, Y) = X - g(Y)$).

Since $x^d \neq D_{d,m}(x)$ for $m \neq 0$ (as $(x + m/x)^d \neq x^d + m^d/x^d$) for $d \geq 2$, Theorem 1 implies there is a finite number of d-th powers in any generalized Fibonacci sequence. Note that the theorem does not give a uniform bound for all $d \geq 0$, and hence does not generalize [Bugeaud et al. 2006]. Finally, we note that the proof of the theorem is effective, generalizes to number fields and to other recurrence sequences; see Remark 6.

2. Preliminaries

2.1. ***Chebyshev and Dickson polynomials.*** Recall that $T_d \in \mathbb{Z}[x]$ denotes the normalized[1] Chebyshev polynomial, that is, the unique degree-d polynomial which satisfies $T_d(x + 1/x) = x^d + 1/x^d$. For an algebraic $m \in \mathbb{C} \setminus \{0\}$, the Dickson polynomial $D_{d,m} \in \mathbb{Q}(m)[x]$ satisfies $D_{d,m}(x + m/x) = x^d + m^d/x^d$ [Schinzel

[1]The usual Chebychev polynomials are $\frac{1}{2}T_d(2x)$; see [Zieve and Mueller 2008, §3].

2000, §1.4, Corollary 2]. Thus

$$D_{d,m}\left(x+\frac{m}{x}\right)=x^d+\frac{m^d}{x^d}=m^{d/2}\left(\frac{x^d}{m^{d/2}}+\frac{m^{d/2}}{x^d}\right)$$
$$=m^{d/2}T_d\left(\frac{x}{\sqrt{m}}+\frac{\sqrt{m}}{x}\right)=m^{d/2}T_d\left(\frac{1}{\sqrt{m}}\left(x+\frac{m}{x}\right)\right). \quad (1)$$

Two polynomials f, g are called equivalent if $f=\ell_1\circ g\circ\ell_2$ for some linear $\ell_1,\ell_2\in\mathbb{C}[x]$. We use the following well known fact [Schinzel 2000, §1.4, Lemma 4]:

Lemma 2. *Suppose $g(x)\in\mathbb{C}[x]$ is of degree $d>1$ and satisfies*

$$(g(x)-a_1)(g(x)-a_2)=(x-r_1)(x-r_2)R(x)^2$$

for complex $a_1\neq a_2$, $r_1\neq r_2$, and $R(x)\in\mathbb{C}[x]$. Then $g(x)=\ell_1\circ T_d\circ\ell_2$, where

$$\ell_1(x)=\pm\frac{a_1-a_2}{4}x+\frac{a_1+a_2}{2}\quad\text{and}\quad\ell_2^{-1}(x)=\frac{r_1-r_2}{4}x+\frac{r_1+r_2}{2}.$$

Similar results to the following lemma and corollary are well known in particular cases, including $p(x)=T_d(x)$ [Lidl et al. 1993, Lemma 6.15]. Since the proof appears to be new, we give it here:

Lemma 3. *Let $K\subset\mathbb{C}$ be a subfield, $\varepsilon\in\mathbb{C}^\times$, and $\mu(x)\in\mathbb{C}[x]\setminus K[x]$ be of degree* 1. *Suppose that all roots of $p(x)\in K[x]$ are real, and the minimal root r is different from the maximal one R. If $\varepsilon p(\mu(x))\in K[x]$, then $R+r\in K$ and $\mu(x)=\sqrt{m}\eta(x)+(R+r)/2$ for linear $\eta\in K[x]$, and $m\in K^\times$. Furthermore, $\varepsilon\in\sqrt{m}\cdot K$ if $p(x-(R+r)/2)$ is an odd polynomial, and $\varepsilon\in K$ otherwise.*

Proof. Let A be the set of roots of $p(x)$. Notice that $\mu^{-1}(A)$ is the set of zeros of $g(x):=\varepsilon p(\mu(x))\in K[x]$. Since $p(x),g(x)\in K[x]$, both A and $\mu^{-1}(A)$ are invariant under every K-automorphism σ, giving

$$\mu^{-1}(A)=\sigma(\mu^{-1}(A))=\sigma(\mu^{-1})(\sigma(A))=\sigma(\mu^{-1})(A).$$

Write $\mu^{-1}(x)=\gamma x+\delta$, so that $\sigma(\mu^{-1})(x)=\sigma(\gamma)x+\sigma(\delta)$. Note that the points of A lie on a segment whose endpoints are R, r. Thus the points $\mu^{-1}(A)$ lie on a segment in $\mathbb{C}$ whose end points are either $(\mu^{-1}(R),\mu^{-1}(r))=(\sigma(\mu^{-1})(R),\sigma(\mu^{-1})(r))$ or $(\mu^{-1}(R),\mu^{-1}(r))=(\sigma(\mu^{-1})(r),\sigma(\mu^{-1})(R))$. In the former case

$$\begin{aligned}R\gamma+\delta&=R\sigma(\gamma)+\sigma(\delta),\\ r\gamma+\delta&=r\sigma(\gamma)+\sigma(\delta),\end{aligned} \quad (2)$$

while in the latter, the ends flip:

$$\begin{aligned}R\gamma+\delta&=r\sigma(\gamma)+\sigma(\delta),\\ r\gamma+\delta&=R\sigma(\gamma)+\sigma(\delta).\end{aligned} \quad (3)$$

Since $R \neq r$, for σ in case (2), we have $\sigma(\delta) = \delta$ and $\sigma(\gamma) = \gamma$. For σ in case (3), σ flips the ends, and hence σ^2 fixes them, so that σ^2 is in case (2), that is $\sigma^2(\delta) = \delta$. Moreover, taking the difference of the two equations in (3) gives $\sigma(\gamma) = -\gamma$. Since $\sigma(\gamma) = \pm\gamma$ for every K-automorphism σ, we have $\gamma = \sqrt{m}$ for $m \in K$. Plugging this into (3) gives $(R+r)\sqrt{m} = \sigma(\delta) - \delta$.

Since (2) and (3) imply that every K-automorphism that fixes $\gamma = \sqrt{m}$ also fixes δ, we have $\delta = a + b\sqrt{m} \in K[\sqrt{m}]$. If $R = -r$, for σ as in (3) we also have $\sigma(\delta) = \delta$, and hence $\delta \in K$. Otherwise, case (3) gives $(R+r)\gamma = 2b\gamma$, and hence

$$b = \frac{R+r}{2} \in K$$

and

$$\mu^{-1}(x) = \sqrt{m}\left(x + \frac{R+r}{2}\right) + a \quad \text{for } a \in K.$$

Thus

$$\mu(x) = \frac{1}{\sqrt{m}}(x-a) - \frac{R+r}{2}.$$

Note that since μ is assumed not to be in $K[x]$, case (2) holds for some σ and hence $R + r \in K$.

Write

$$q(x) := p\left(x - \frac{R+r}{2}\right) \in K[x],$$

and

$$\mu(x) + \frac{R+r}{2} = \sqrt{m}\eta(x)$$

for a linear $\eta \in K[x]$, so that $\varepsilon p(\mu(x)) = \varepsilon q(\sqrt{m}\eta(x))$ and hence $\varepsilon q(\sqrt{m}x)$ are in $K[x]$. If q is odd, then $q(\sqrt{m}x) = \sqrt{m}s(x)$, where $s(x) \in K[x]$; thus $\varepsilon \cdot \sqrt{m} \in K$. Otherwise, q has at least one nonzero monomial of even degree, say $\varepsilon q_{2k} m^k x^{2k} \in K \cdot x^{2k}$ with $q_{2k} \in K \setminus \{0\}$, so that $\varepsilon \in K$. □

Corollary 4. *Let $K \subseteq \mathbb{C}$ be a field and $d \geq 2$ an integer. Suppose $q(x) := \varepsilon T_d(\mu(x)) \in K[x]$ for $\varepsilon \in \mathbb{C}^\times$ and linear $\mu \in \mathbb{C}[x]$. Then $q(x) = \varepsilon m^{d/2} D_{d,m}(m\eta(x))$ for $m \in K^\times$ such that $\varepsilon m^{d/2} \in K$, where $\eta(x) = \mu(x)/\sqrt{m} \in K[x]$.*

Proof. All roots of $T_d(x)$ are real, the maximal root is $R = 2$, and the minimal root is $r = -2$. If $\mu \in K[x]$, then $\varepsilon \in K$, and the claim follows with $m = 1$. Otherwise, as $R + r = 0$, Lemma 3 yields that $\mu(x) = \sqrt{m}\eta(x) \in K[x]$ for some $m \in K^\times$. Since in addition $T_d(x) = m^{-d/2} D_{d,m}(\sqrt{m}x)$ by (1), we get $q(x) = \varepsilon m^{d/2} D_{d,m}(m\eta(x))$. □

2.2. ***Siegel's theorem.*** We shall also use the following explicit version of Siegel's theorem [LeVeque 1964, Theorem 1] for the special case of hyperelliptic curves:

Theorem 5. *Let $f(x) \in K[x]$ be a polynomial of degree ≥ 3 over a number field K which factors as $f(x) = a\prod_{i=1}^{s}(x-\alpha_i)^{r_i}$ for distinct $\alpha_i \in \mathbb{C}$, for $a \in K$, and integers $r_i \geq 1$. If $y^2 = f(x)$ has infinitely many solutions in the ring of integers of K, then at most two of the r_i are odd.*

In the case $K = \mathbb{Q}$ and $u(x) = \sum_{i=0}^{n} a_i x^i \in \mathbb{Q}[x]$ has at least three distinct simple roots, Baker [1969, Theorem 2][2] gave an (exponential) bound on $\max\{|x|, |z|\}$ for solutions (x, z) to $z^2 = u(x)$, in terms of $\max\{|a_i| : i = 0, \ldots, n\}$. In other words, the solution to $z^2 = u(x)$ is effective.

An effective solution to $y^2 = f(x)$ can be deduced in the case at least three of the multiplicities r_i of roots of f are odd. Indeed, this case reduces to the situation in Baker's theorem by a variable change $z = y/h(x)^2$, where $h(x) \in \mathbb{Q}[x]$ satisfies $h(x)^2 \mid f(x)$ and $f(x)/h(x)^2$ is square-free, since the resulting equation is $z^2 = u(x)$ with $u(x) = f(x)/h(x)^2$ has at least three simple roots.

3. Proof of Theorem 1

Recall that to a generalized Fibonacci sequence $F^{a,b}$, we attach $\chi_{a,b} = a^2 + ab - b^2$. Note that $\chi_{a,b} \neq 0$ for $(a, b) \in \mathbb{Z}^2 \setminus \{(0, 0)\}$. Indeed, if $a^2 + ab - b^2 = 0$, then

$$b = \frac{a \pm \sqrt{5a^2}}{2} = a\left(\frac{1 \pm \sqrt{5}}{2}\right).$$

Since $a, b \in \mathbb{Z}$, this implies $\chi_{a,b} = 0$ if and only if $a = b = 0$.

Proof of Theorem 1. If $a = b = 0$, the only element in the sequence is 0, and the theorem holds trivially. Henceforth, as above, we may assume $\chi_{a,b} \neq 0$. We first claim that each element of $F^{a,b}$ gives rise to a unique integer solution to one of the two equations $y^2 = 5x^2 \pm 4\chi_{a,b}$ given by $(x, y) = (F_n, \sqrt{5F_n^2 - 4\chi_{a,b}})$ for even n, and $(x, y) = (F_n, \sqrt{5F_n^2 + 4\chi_{a,b}})$ for odd n. This is a consequence of the following generalized Cassini identity. By the definition of F_n, we have

$$\begin{pmatrix} F_{n+2} & F_{n+1} \\ F_{n+1} & F_n \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} F_{n+1} & F_n \\ F_n & F_{n-1} \end{pmatrix}$$

and hence inductively one has

$$\begin{pmatrix} F_{n+2} & F_{n+1} \\ F_{n+1} & F_n \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}^n \begin{pmatrix} a+b & b \\ b & a \end{pmatrix}.$$

Taking the determinants of both sides gives the generalized Cassini identity:

$$F_{n+2}F_n - F_{n+1}^2 = (-1)^n(a^2 + ab - b^2).$$

Substituting $F_{n+2} = F_n + F_{n+1}$ into the equation, we have the quadratic equation in F_{n+1}

$$F_n^2 + F_nF_{n+1} - F_{n+1}^2 = (-1)^n\chi_{a,b},$$

and hence

$$F_{n+1} = \frac{F_n \pm \sqrt{5F_n^2 + (-1)^{n+1}4\chi_{a,b}}}{2}.$$

[2] Since then considerably better bounds have been found.

It follows that $5F_n^2 + 4(-1)^{n+1}\chi_{a,b} = y^2$ for some $y \in \mathbb{Z}$, proving the claim.

Hence, to prove the theorem it suffices to show that if one of the equations

$$y^2 = 5g(x)^2 + 4\chi, \quad \text{with } \chi = \chi_{a,b} \text{ or } -\chi_{a,b}, \tag{4}$$

has infinitely many solutions $(x, y) \in \mathbb{Z}^2$, then $g(x) = \alpha_{\chi,d,m} D_{d,m}(\mu(x))$ for some linear $\mu \in \mathbb{Q}[x]$ and some $m \in \mathbb{Z}$, with $d = \deg g$. Without loss of generality assume the equation with $\chi = \chi_{a,b}$ has infinitely many solutions.

Set $f(x) := 5g(x)^2 + 4\chi$, $d := \deg(g)$, and write $f(x) = \alpha \prod_{i=1}^{u} (x - x_i)^{r_i} \in \mathbb{C}[x]$. As $\deg f \geq 4$, if $y^2 = f(x)$ has infinitely many solutions $(x, y) \in \mathbb{Z}^2$, then Theorem 5 implies that at most two of the r_i are odd.

The case where exactly one of the r_i is odd, contradicts $\deg f = 2d$ is even. In the case all r_i are even, $f(x) = h(x)^2$ for some $h \in \mathbb{C}[x]$. As $\chi \neq 0$, we have

$$\deg(f - g^2) = \deg(h^2 - 5g^2) = \deg(h - \sqrt{5}g) + \deg(h + \sqrt{5}g) = 0,$$

and so $\deg(h \pm \sqrt{5}g) = 0$. This shows that g is constant contradicting $d > 1$.

If exactly two of the r_i are odd, say r_1, r_2, we have the decomposition

$$\left(g(x) - 2\sqrt{-\frac{\chi}{5}}\right)\left(g(x) + 2\sqrt{-\frac{\chi}{5}}\right) = \frac{f(x)}{5} = (x - x_1)(x - x_2)R(x)^2$$

for $R(x) \in \mathbb{C}[x]$, and hence Lemma 2 with $a_1 = -a_2 = 2\sqrt{-\chi/5}$ implies

$$g(x) = \pm\sqrt{-\frac{\chi}{5}}\, T_d(\ell_2(x))$$

for some linear $\ell_2 \in \mathbb{C}[x]$.

Corollary 4 then implies $\ell_2(x) = (1/\sqrt{m})\mu(x)$ for some $\mu \in \mathbb{Q}[x]$, and $m \in \mathbb{Z}\setminus\{0\}$. Furthermore, it shows

$$g(x) = \pm\sqrt{-\frac{\chi}{5m^d}}\, D_{d,m}(\mu(x))$$

with

$$\alpha_{\chi,d,m} = \pm\sqrt{-\frac{\chi}{5m^d}} \in \mathbb{Q}.$$

Finally, to see that $g(x) = x^2$ is never of the above form, note that 0 is a branch point of $g(x)$, that is, the value of g at a root of $g'(x)$. On the other hand, 0 is never a branch point of $\alpha_{\chi,2,m} D_{2,m}(x)$. □

Remark 6. (1) The proof shows that if $g(\mathbb{Z}) \cap F^{a,b}$ is infinite and $d = \deg g$ is even, then $\chi = \chi_{a,b} < 0$. Together with the above observation that $F_n \in g(\mathbb{Z})$ gives a solution to (4) with $\chi = \chi_{a,b}$ for odd n, and with $\chi = -\chi_{a,b}$ for even n, this shows that $F^{a,b}$ intersects $g(\mathbb{Z})$ infinitely often at only one residue class of n mod 2.

(2) In the case $g(\mathbb{Z}) \cap F^{a,b}$ is finite, Theorem 1 is effective, that is, gives an (exponential) bound on the largest value of n for which $F_n \in g(\mathbb{Z})$, in terms of the coefficients of g and $\chi_{a,b}$. Indeed, since the coefficients of f are bounded in

terms of the coefficients of g and $\chi_{a,b}$, and since Siegel's theorem is applied to the hyperelliptic curve $y^2 = f(x)$ in the case $f(x)$ has at least three roots with odd multiplicity, Baker's theorem gives the desired bound as described in Section 2.2.

(3) Finally the proof extends to the recurrence relations $G_{n+2} = \pm G_n + BG_{n+1}$ for rings of integers $R = O_K$ of number fields K, and $B \in R$. The main modification is replacing 5 by $B^2 + 4u$.

Theorem 7. *Let $R = O_K$ be the ring of integers of a number field K and $G^{a,b}$, $a, b \in R$, be the sequence given by $G_0 = a$, $G_1 = b$, and $G_{n+2} = uG_n + BG_{n+1}$ for $u \in \{1, -1\}$, and $B \in R$. Assume $B^2 + 4u \notin R^2$, and set $\chi_G := ua^2 + Bab - b^2$. Let $g(x) \in K[x]$ be a polynomial of degree $d > 1$ different from $\pm\alpha_{\chi,d,m} D_{d,m}(\mu(x))$ for all linear $\mu(x) \in K[x]$, $m \in R$ and*

$$\alpha_{\chi,d,m} = \sqrt{-\frac{\chi \cdot m^d}{B^2 + 4u}},$$

where $\chi = -\chi_G$ if $u = -1$ and $\chi \in \{\chi_G, -\chi_G\}$ if $u = 1$. Then $\#(g(R) \cap G^{a,b}) < \infty$.

Proof. First note that $\chi_G \neq 0$ since

$$b \neq \frac{B \pm \sqrt{B^2 + 4u}}{2} \cdot a,$$

as $B^2 + 4u \notin R^2$.

We adjust the relevant parts from the proof of the main theorem. In the generalized Cassini identity, we have the equation

$$\begin{pmatrix} G_{n+2} & G_{n+1} \\ G_{n+1} & G_n \end{pmatrix} = \begin{pmatrix} B & u \\ 1 & 0 \end{pmatrix} \begin{pmatrix} G_{n+1} & G_n \\ G_n & G_{n-1} \end{pmatrix}$$

and by induction

$$\begin{pmatrix} G_{n+2} & G_{n+1} \\ G_{n+1} & G_n \end{pmatrix} = \begin{pmatrix} B & u \\ 1 & 0 \end{pmatrix}^n \begin{pmatrix} Bb+ua & b \\ b & a \end{pmatrix}.$$

Thus, by taking determinants we find that

$$G_{n+2}G_n - G_{n+1}^2 = (-u)^n(ua^2 + Bab - b^2).$$

Setting $\chi_G = ua^2 + Bab - b^2$, we get the equation

$$uG_n^2 + BG_nG_{n+1} - G_{n+1}^2 = (-u)^n \chi_G,$$

and hence

$$G_{n+1} = \frac{BG_n \pm \sqrt{B^2G_n^2 + 4uG_n^2 - 4(-u)^n \chi_G}}{2}.$$

Thus $(B^2+4u)G_n^2 - 4(-u)^n\chi_G = y^2,\ y \in R$, yielding the analogous equation to (4):

$$y^2 = f(x) = (B^2+4u)g(x)^2 + 4\chi,$$

where $\chi = -\chi_G$ when $u = -1$ and it takes two possible values $\chi \in \{\chi_G, -\chi_G\}$ when $u = 1$. Similarly to the proof of Theorem 1, an application of Theorem 5 shows that there are finitely many solutions in O_K to the equation $y^2 = f(x)$ if $f(x)$ has at least three roots of odd multiplicity. Again, it cannot have a single root of odd multiplicity since f is of even degree, and $f(x)$ cannot be a square since then $f(x) - (\sqrt{B^2+4u}g(x))^2 = 4\chi \neq 0$ is a nonzero constant difference of squares in $\mathbb{C}[x] \setminus \mathbb{C}$. Thus

$$f(x) = (x-x_1)(x-x_2)R^2(x)$$

for distinct $x_1, x_2 \in \mathbb{C}$ and $R(x) \in \mathbb{C}[x]$. Lemma 2 applied to the decomposition

$$\left(g(x) - 2\sqrt{-\frac{\chi}{B^2+4u}}\right)\left(g(x) + 2\sqrt{-\frac{\chi}{B^2+4u}}\right) = \frac{f(x)}{B^2+4u},$$

gives

$$g(x) = \pm\sqrt{-\frac{\chi}{B^2+4u}}\,T_d(\ell_2(x)).$$

The result now follows from Corollary 4. □

Acknowledgements

The authors thank the Mentoring Program of the Szold Institute for the opportunity it gave Elrazik and Ostrov in entering academic research. We thank Umberto Zannier, and Michael Zieve, and the referees for helpful comments. This paper originated as a project of Ostrov as part of the program, advised by Berman and Neftin, with the help of Elrazik. Ostrov was supported by the Israel Science Foundation (grant no. 662/15), and Neftin by the Binational Science Foundation (grant no. 2014173).

References

[Baker 1969] A. Baker, "Bounds for the solutions of the hyperelliptic equation", *Proc. Cambridge Philos. Soc.* **65** (1969), 439–444. MR Zbl

[Bugeaud et al. 2006] Y. Bugeaud, M. Mignotte, and S. Siksek, "Classical and modular approaches to exponential Diophantine equations, I: Fibonacci and Lucas perfect powers", *Ann. of Math.* (2) **163**:3 (2006), 969–1018. MR Zbl

[Cohn 1964] J. H. E. Cohn, "Square Fibonacci numbers, etc", *Fibonacci Quart.* **2**:2 (1964), 109–113. MR Zbl

[Corvaja and Zannier 1998] P. Corvaja and U. Zannier, "Diophantine equations with power sums and universal Hilbert sets", *Indag. Math.* (*N.S.*) **9**:3 (1998), 317–332. MR Zbl

[LeVeque 1964] W. J. LeVeque, "On the equation $y^m = f(x)$", *Acta Arith.* **9** (1964), 209–219. MR Zbl

[Lidl et al. 1993] R. Lidl, G. L. Mullen, and G. Turnwald, *Dickson polynomials*, Pitman Monographs and Surveys in Pure and Applied Mathematics **65**, Longman Scientific & Technical, Harlow, 1993. MR Zbl

[Nagy et al. 2019] M. Nagy, S. R. Cowell, and V. Beiu, "Survey of cubic Fibonacci identities — when cuboids carry weight", preprint, 2019. arXiv

[Schinzel 2000] A. Schinzel, *Polynomials with special regard to reducibility*, Encyclopedia of Mathematics and its Applications **77**, Cambridge University Press, 2000. MR Zbl

[Zieve and Mueller 2008] M. E. Zieve and P. Mueller, "On Ritt's polynomial decomposition theorems", preprint, 2008. arXiv

Received: 2019-12-14 Revised: 2020-04-21 Accepted: 2020-07-26

phzheveuubyhy@gmail.com *Department of Mathematics, Technion – Israel Institute of Technology, Haifa, Israel*

dneftin@technion.ac.il *Department of Mathematics, Technion – Israel Institute of Technology, Haifa, Israel*

berman@technion.ac.il *Department of Mathematics, Technion – Israel Institute of Technology, Haifa, Israel*

reyad@campus.technion.ac.il *Department of Mathematics, Technion – Israel Institute of Technology, Haifa, Israel*

INVOLVE 13:4 (2020)
https://doi.org/10.2140/involve.2020.13.607

Stability and asymptotic analysis of the Föllmer–Schweizer decomposition on a finite probability space

Sarah Boese, Tracy Cui, Samuel Johnston,
Sylvie Vega-Molino and Oleksii Mostovyi

(Communicated by Jonathon Peterson)

First, we consider the problem of hedging in complete binomial models. Using the discrete-time Föllmer–Schweizer decomposition, we demonstrate the equivalence of the backward induction and sequential regression approaches. Second, in incomplete trinomial models, we examine the extension of the sequential regression approach for approximation of contingent claims. Then, on a finite probability space, we investigate stability of the discrete-time Föllmer–Schweizer decomposition with respect to perturbations of the stock price dynamics and, finally, perform its asymptotic analysis under simultaneous perturbations of the drift and volatility of the underlying discounted stock price process, where we prove stability and obtain explicit formulas for the leading-order correction terms.

1. Introduction

In practice, financial models are not exact — as in any field, modeling based on real data introduces some degree of error. Therefore, it is important to understand the effect error has on the calculations and assumptions we make on the model. In this paper, we focus on the stability and asymptotic analysis of the Föllmer–Schweizer decomposition, as among the pricing and hedging approaches, in incomplete markets, it gives the best approximation of a given contingent claim in the sense of the least-squares error, which is one of the most natural criteria used in practice. Further, in complete models, the Föllmer–Schweizer decomposition is consistent with the backward induction, which is another canonical method in financial mathematics.

MSC2010: primary 60G07, 93E20, 91G10, 91G20, 90C31; secondary 60H30, 93E24.
Keywords: Föllmer–Schweizer decomposition, simultaneous perturbations of the drift and volatility, asymptotic analysis, stability.
This paper is a part of an REU project conduced in Summer 2019 at the University of Connecticut. The project was supported by the National Science Foundation under grants DMS-1659643 and DMS-1848339.

Steven E. Shreve [2004a] introduces option pricing in a highly accessible manner. His text predominantly focuses on the binomial model, and in this paper, we go beyond it, as there are many models used in practice that are not binomial. As the most natural extension of a binomial model is a trinomial one, below, we also give it special consideration. Note that both binomial and trinomial models, despite their simplicity, are widely used in approximations of pricing and hedging in more advanced models, including the continuous-time ones; see, e.g., [Brigo and Mercurio 2006].

In this paper, we extend the introduction to asset hedging given by Shreve to the strategy of sequential regression, keeping the discrete-time framework but allowing for consideration of other market models, including incomplete ones. In the complete case, we show that the strategy of sequential regression introduced by Föllmer and Schweizer [1989] is equivalent to Shreve's recursive hedging formula. We then extend our discussion to the incomplete trinomial model, after which we show small market perturbations have a little effect, which we quantify, on hedging strategies and option pricing, and derive formulas for correction factors.

The remainder of this paper is organized as follows. In Section 2, we formulate the minimization problem and define the Föllmer–Schweizer decomposition. Section 3 contains its investigation in complete binomial markets, where we also prove the equivalence of the approach based on the Föllmer–Schweizer decomposition to the backward induction. Section 4 presents the discussion of the general incomplete case. In Section 5, we revisit the stability question in the context of perturbations of the model parameters, where we introduce a parametrization of perturbations that allows for simultaneous distortions of its drift and volatility of the underlying stock price process. We prove the stability of the Föllmer–Schweizer decomposition under such perturbations. Finally, in Section 6, we obtain explicit formulas for the first-order correction terms of each component of the decomposition under such perturbations, including the correction to the optimal trading strategy.

2. The discrete-time Föllmer–Schweizer decomposition

Let $(\Omega, \mathbb{P})$ be a finite probability space, N a fixed positive integer and $\mathbb{F} = (\mathcal{F}_n)_{n=0,1,\dots,N}$ a filtration, i.e., an increasing family of subalgebras, each containing Ω and $\varnothing$. Assume that $\mathcal{F}_0$ is trivial and $\mathcal{F}_N$ contains all subsets of Ω. As we work on a finite probability space, without loss of generality, we suppose that $\mathbb{P}[\omega] > 0$ for every $\omega \in \Omega$. We suppose that there is a bank account, which we will use as a numéraire, and in particular, that its price process is equal to 1 at all times. Let $S = (S_n)_{n=0,1,\dots,N}$ be a real-valued, $\mathbb{F}$-adapted process; i.e., each S_n is $\mathcal{F}_n$-measurable. S describes the discounted price process of a stock. We denote by

$$\Delta S_n := S_n - S_{n-1} \quad \text{for } n = 1, \dots, N$$

the increments of S. We call a process $\vartheta = (\vartheta_n)_{n=1,\dots,N}$ *predictable* if ϑ_n is $\mathcal{F}_{n-1}$-measurable for each n. Let Θ be the set of all predictable processes ϑ that financially correspond to self-financing trading strategies, in view of the presence of money market account, which we use as a numéraire.

Definition 2.1. For $\vartheta \in \Theta$, we define the process

$$G_n(\vartheta) := \sum_{j=1}^{n} \vartheta_j \Delta S_j \quad \text{for } n = 0, 1, \dots, N.$$

For a given a random variable V_N and $c \in \mathbb{R}$, one can consider the following problem posed in [Schweizer 1995]:

$$\text{minimize } \mathbb{E}[(V_N - c - G_N(\vartheta))^2] \text{ over all } \vartheta \in \Theta \text{ and } c \in \mathbb{R}. \tag{1}$$

Interpretation 2.2. As we view S_n as the price at time n of a risky financial asset, the process ϑ describes the trading strategy of some investor in the market, where ϑ_n is the number of shares of stock held between the times n and $n+1$. The process $G(\vartheta)$ becomes the *gains from the trade* process. We now interpret V_N as a *nontraded security* measured in the units of the bank account with maturity N and c as the *initial capital*. Thus problem (1) can be interpreted as finding a self-financing trading strategy that gives the best least-squares approximation of V_N. Mathematically, (1) is also closely related to the problem considered in [Schweizer 1994], finding the best approximation of a random variable by a stochastic integral (plus a constant). Quadratic optimization problems of the form (1) also appear in the asymptotic analysis of stochastic control problems with respect to perturbations of the initial data, where they govern the second-order correction terms; see [Kramkov and Sîrbu 2006a; 2006b, Mostovyi and Sîrbu 2019; Mostovyi 2020] for details.

A solution (1), given in terms of an explicit formula for an optimal trading strategy $\hat{\vartheta}$, is known as *sequential regression* and is shown in [Föllmer and Schweizer 1989]. For a general probability space, such a solution is subject to additional conditions on S and is closely related to the discrete-time *Föllmer–Schweizer decomposition*, defined below.

Definition 2.3. We use the definition of the *nondegeneracy condition* (ND) as given in [Schweizer 1995]; that is, S satisfies (ND) if there exists a constant $\delta \in (0, 1)$ such that

$$(\mathbb{E}[\Delta S_n \mid \mathcal{F}_{n-1}])^2 \le \delta \mathbb{E}[\Delta S_n^2 \mid \mathcal{F}_{n-1}] \quad \mathbb{P}\text{-a.s. for } n = 1, \dots, N.$$

Remark 2.4. We note that on finite probability spaces, (ND) holds in nontrivial (or rather nondegenerate) cases.

Definition 2.5. We now introduce the *discrete Föllmer–Schweizer decomposition*, following [Schweizer 1995]. Let

$$S = M + A$$

be the semimartingale decomposition of S into a martingale M and a predictable process A. The random variable V_N admits the *discrete Föllmer–Schweizer decomposition* if it can be written as

$$V_N = V_0 + \sum_{j=1}^{N} \hat{\vartheta}_j \Delta S_j + L_N \tag{2}$$

for some $V_0 \in \mathbb{R}$, a process $\hat{\vartheta} \in \Theta$, and a $\mathbb{P}$-martingale L, such that

(i) L and M are orthogonal, i.e., LM is a $\mathbb{P}$-martingale, and

(ii) $\mathbb{E}[L_0] = 0$.

Note that when $\mathcal{F}_0$ is trivial, the latter condition reads $L_0 = 0$.

Using the sequential regression approach, following [Föllmer and Schweizer 1989], we obtain the following formula for an optimal hedging strategy $\hat{\vartheta}$:

$$\hat{\vartheta}_n := \frac{\mathrm{Cov}_{\mathcal{F}_{n-1}}\left[V_N - \sum_{j=n+1}^{N} \hat{\vartheta}_j \Delta S_j, \Delta S_n\right]}{\mathrm{Var}_{\mathcal{F}_{n-1}}[\Delta S_n]}, \quad n = 1, \ldots, N, \tag{3}$$

where $\mathrm{Cov}_{\mathcal{F}_{n-1}}[\,\cdot\,,\cdot\,]$ and $\mathrm{Var}_{\mathcal{F}_{n-1}}[\,\cdot\,]$ denote the conditional covariance and variance, respectively. This demonstrates the richness of the FS-decomposition as an analytic tool. With very limited assumptions about V_N, we are able to obtain an explicit formula for the optimal (in the sense of (1)) hedging strategy. Furthermore, this hedging formula holds in both complete and incomplete markets, which are discussed in the following sections.

3. Complete markets

In the settings of the previous section, when the market is complete, every contingent claim V_N can be represented as

$$V_N = V_0 + \sum_{j=1}^{N} \vartheta_j \Delta S_j$$

for some $\vartheta \in \Theta$ and $V_0 \in \mathbb{R}$. Note that this situation corresponds to $L = 0$ in Definition 2.5. Put differently, V_N can now be obtained by trading between the money market account and the stock.

Binomial asset pricing model. We now introduce a simple framework for the problem following [Shreve 2004a], however directly using the bank account as a

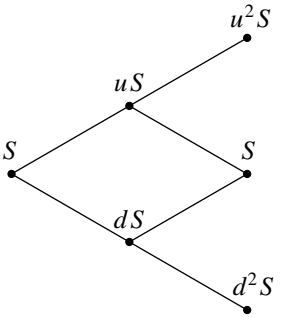


Figure 1. Example of a 2-period binomial model.

numéraire. Consider a *binomial asset pricing model*, where at each time step k, S_{k+1} can take one of two values, uS_k or dS_k, with probabilities $p \in (0, 1)$ and $q := 1 - p$, and where $u > 1$ and $d \in (0, 1)$, known as the *up factor* and *down factor*, respectively, are fixed positive constants with $u > d$. The value at each time k is determined by a (not necessarily fair) coin flip ω_k, which can take either the value H or T and is independent of other coin tosses. Let $\mathbb{F} = (\mathcal{F}_n)_{n=0,\dots,N}$ be the filtration, where each $\mathcal{F}_n$ contains information about the first n coin tosses. An example of a 2-period binomial model is shown in Figure 1.

Remark 3.1. For Figure 1, for the binomial model, we considered the case when $d = 1/u$. That is, after an even number of time steps, the stock price returns to its original value if we flip exactly the same number of heads and tails. However, in general, it is not necessarily the case that $u = 1/d$.

Let X be the replicating wealth process for the contingent claim V_N, i.e., a self-financing process starting from the initial wealth X_0 and such that

$$X_N = V_N.$$

Classical backward induction approach, for which we refer to [Shreve 2004a], assures that the number of shares of stock in the replicating portfolio for V_N held between times n and $n + 1$, can be obtained via the formula

$$\vartheta_{n+1}(\omega) = \frac{X_{n+1}(\omega H) - X_{n+1}(\omega T)}{S_{n+1}(\omega H) - S_{n+1}(\omega T)}, \tag{4}$$

where $\omega = \omega_1, \dots, \omega_n$. This result is known as the delta-hedging rule; see [Shreve 2004a, Theorem 1.2.2, p. 12].

The main result of this section is proving that in complete binomial settings the delta-hedging rule gives the same strategy as the Föllmer–Schweizer decomposition.

Proposition 3.2. *In complete binomial settings, formulas* (3) *and* (4) *are equivalent.*

Proof. Below $\hat{\vartheta}$ will denote the strategy obtained from formula (3), and ϑ will denote the replicating strategy given by (4) for each n. First, for $n = N$, (3) reads

$$\hat{\vartheta}_N = \frac{\mathrm{Cov}_{\mathcal{F}_{N-1}}[V_N, \Delta S_N]}{\mathrm{Var}_{\mathcal{F}_{N-1}}[\Delta S_N]} = \frac{\mathrm{Cov}_{\mathcal{F}_{N-1}}[X_N, \Delta S_N]}{\mathrm{Var}_{\mathcal{F}_{N-1}}[\Delta S_N]}, \tag{5}$$

as $X_N = V_N$. Also, since[1]

$$X_{n+1} = X_n + \vartheta_{n+1}(S_{n+1} - S_n), \quad n = 0, \ldots, N-1,$$

we have

$$X_N - \mathbb{E}_{\mathcal{F}_{N-1}}[X_N] = \vartheta_N(\Delta S_N - \mathbb{E}_{\mathcal{F}_{N-1}}[\Delta S_N]).$$

Therefore, we can rewrite $\mathrm{Cov}_{\mathcal{F}_{N-1}}[V_N, \Delta S_N]$ as

$$\begin{aligned}\mathrm{Cov}_{\mathcal{F}_{N-1}}[V_N, \Delta S_N] &= \mathbb{E}_{\mathcal{F}_{N-1}}[(X_N - \mathbb{E}_{\mathcal{F}_{N-1}}[X_N])(\Delta S_N - \mathbb{E}_{\mathcal{F}_{N-1}}[\Delta S_N])] \\ &= \vartheta_N \mathbb{E}_{\mathcal{F}_{N-1}}[(\Delta S_N - \mathbb{E}_{\mathcal{F}_{N-1}}[\Delta S_N])(\Delta S_N - \mathbb{E}_{\mathcal{F}_{N-1}}[\Delta S_N])] \\ &= \vartheta_N \,\mathrm{Var}_{\mathcal{F}_{N-1}}[\Delta S_N].\end{aligned}$$

And thus

$$\vartheta_N = \frac{\mathrm{Cov}_{\mathcal{F}_{N-1}}[V_N, \Delta S_N]}{\mathrm{Var}_{\mathcal{F}_{N-1}}[\Delta S_N]}.$$

Comparing ϑ_N to $\hat{\vartheta}_N$ from (5), we deduce that they coincide. This also implies

$$X_{N-1} = X_N - \hat{\vartheta}_N \Delta S_N.$$

Or equivalently, we have

$$X_N - \sum_{j=N}^{N} \hat{\vartheta}_N \Delta S_N = X_{N-1}. \tag{6}$$

The latter expression, however, is exactly the term appearing in $\hat{\vartheta}_{N-1}$ in (3), which via (6), we can rewrite as

$$\hat{\vartheta}_{N-1} = \frac{\mathrm{Cov}_{\mathcal{F}_{N-2}}\left[X_N - \sum_{j=N}^{N} \hat{\vartheta}_N \Delta S_{N-1}, \Delta S_{N-1}\right]}{\mathrm{Var}_{\mathcal{F}_{N-2}}[\Delta S_{N-1}]} = \frac{\mathrm{Cov}_{\mathcal{F}_{N-2}}[X_{N-1}, \Delta S_{N-1}]}{\mathrm{Var}_{\mathcal{F}_{N-2}}[\Delta S_{N-1}]},$$

which differs from (5) only by the value of the index. Therefore, line by line applying the argument above used for proving that $\vartheta_N = \hat{\vartheta}_N$, we can show that $\vartheta_{N-1} = \hat{\vartheta}_{N-1}$. Proceeding in such a way, we can show that $\vartheta_n = \hat{\vartheta}_n$ for each $n \in \{0, \ldots, N\}$. □

Remark 3.3. Notice that Proposition 3.2 only demonstrates the equivalence of the specific strategies of backward recursion and sequential regression in the binomial model, not necessarily the uniqueness of the minimizing strategy. However, the

[1]We use indices for ϑ in a way consistent with [Schweizer 1995], so that ϑ is predictable.

uniqueness of the solution to (1), that is, of the optimal stochastic integral, the associated strategy, and a constant follows from the strict convexity of the quadratic objective in (1) and some computations under the risk-neutral measure.

Example 3.4. Consider a 3-step binomial asset pricing model with $S_0 = 4$, $u = 2$, $d = \frac{1}{2}$, $r = \frac{1}{4}$, $p = \frac{1}{2}$, $q = \frac{1}{2}$, and a *European call option* expiring at time $N = 3$ with strike price $K = 1$. Note that in this market, the one-step risk-neutral probabilities are $\tilde{p} = \tilde{q} = \frac{1}{2}$ and the nonzero interest rate can be handled by considering the discounted stock. We will illustrate the optimal hedge using both (3) and (4). Recall that the value at time N of a European call is given by

$$V_N = (S_N - K)^+. \tag{7}$$

We compute the stock prices and discounted asset values at each time step for each possible combination of coin flips using the formulas

$$S_n(\omega_1, \ldots, \omega_n) = S_0 * u^{(\#\text{ heads})} * d^{(\#\text{ tails})},$$
$$V_n(\omega_1, \ldots, \omega_n) = \left(\frac{1}{1+r}\right)\tilde{\mathbb{E}}[V_{n+1} \mid \omega_1, \ldots, \omega_n],$$

where $\mathbb{P}$ is the risk-neutral probability measure. The trees of stock prices and discounted asset values are shown in Figure 2. For brevity, we will calculate a hedge using both formulas at time 2, since $A_N = V_N$ in the sequential regression formula at the penultimate time step. We aim to show that $\vartheta_2(TT) = \xi_2(TT)$. Calculating $\vartheta_2(TT)$ using (4) yields

$$\frac{1-0}{2-\frac{1}{2}} = \frac{1}{\frac{3}{2}} = \frac{2}{3}.$$

Using (3), $\xi_2(TT)$ is given by

$$\begin{aligned}
\frac{\operatorname{Cov}_{\mathcal{F}_2}[V_3, \Delta S_3 \mid TT]}{\operatorname{Var}_{\mathcal{F}_2}[(\Delta S_3 \mid TT)]} &= \frac{\mathbb{E}_{\mathcal{F}_2}[(V_3 - \mathbb{E}_{\mathcal{F}_2}[V_3 \mid TT])(\Delta S_3 - \mathbb{E}_{\mathcal{F}_2}[\Delta S_3 \mid TT]) \mid TT]}{\mathbb{E}_{\mathcal{F}_2}[(\Delta S_3 - \mathbb{E}_{\mathcal{F}_2}[\Delta S_3 \mid TT])^2 \mid TT]} \\
&= \frac{\mathbb{E}_{\mathcal{F}_2}\left[\left(V_3 - \left(\frac{1}{2} * 1 + \frac{1}{2} * 0\right)\right)\left(\Delta S_3 - \left(\frac{1}{2} * 1 + \frac{1}{2} * -\frac{1}{2}\right)\right) \mid TT\right]}{\mathbb{E}_{\mathcal{F}_2}[\Delta S_3^2 \mid TT] - \mathbb{E}_{\mathcal{F}_2}[\Delta S_3 \mid TT]^2} \\
&= \frac{\mathbb{E}_{\mathcal{F}_2}\left[\left(V_3 - \frac{1}{2}\right)\left(\Delta S_3 - \frac{1}{4}\right) \mid TT\right]}{\frac{5}{8} - \frac{1}{16}} \\
&= \frac{\left(\frac{1}{2}\right)\left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{4}\right) + \left(\frac{1}{2}\right)\left(0 - \frac{1}{2}\right)\left(-\frac{1}{2} - \frac{1}{4}\right)}{\frac{9}{16}} \\
&= \frac{\frac{3}{16} + \frac{3}{16}}{\frac{9}{16}} = \frac{2}{3}.
\end{aligned}$$

This illustrates the equivalence of (3) and (4) in the context of this example.

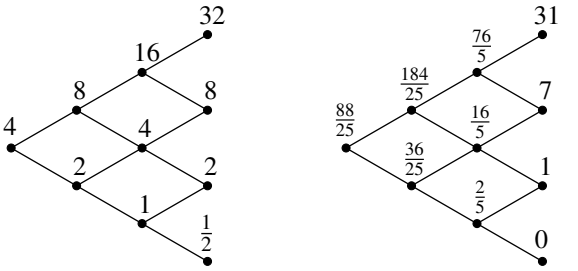


Figure 2. Stock price (left) and option value (right) trees.

Remark 3.5. Although (3) and (4) give us equivalent results in the binomial case, it is important to note that (4) is specifically limited to the binomial case, while (3) can be extended to general discrete-time market models, including incomplete models. In the following section, we apply (3) to the incomplete trinomial model.

4. Incomplete markets

While the binomial model is often used as an introductory tool, most models used in practice exhibit incompleteness, that is, for some securities, it is not possible to construct a hedging strategy that perfectly replicates its payoff. We now introduce a tractable example of an incomplete market.

Trinomial asset pricing model. An example of an incomplete market is based on the *trinomial asset pricing model*. Similar to the binomial model, we have a risky asset and a risk-free asset, and the value of the risky asset at each time step is determined by a small set of outcomes. This time, we have three possible outcomes for the coin flip instead of two. That is, along with the possibility of an increase by a factor of u and decrease by a factor of d, we allow for the possibility that the stock price does not change between two consecutive time steps. See Figure 3.

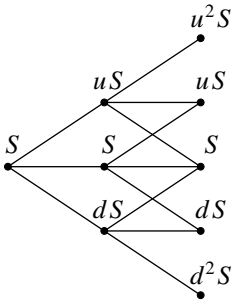


Figure 3. Example of a 2-period trinomial model.

Remark 4.1. For the above formulation of the trinomial model, we again require that $u = 1/d$. However, as in the binomial model, this requirement is not necessary.

Attempting backward recursion on this model using the nondiscounted wealth process, we get

$$X_2(\omega_1\omega_2) = (1+r)(X_1(\omega_1) - \vartheta_1(\omega_1)S_1(\omega_1)) + \vartheta_1(\omega_1)S_2(\omega_1\omega_2),$$
$$X_1(\omega_1) = (1+r)(X_0 - \vartheta_0 S_0) + \vartheta_0 S_1(\omega_1).$$

Note that there are three possible values for ω_1 and three possible values for ω_2, and we must solve for ϑ_0, X_0, each $\vartheta_1(\omega_1)$, and each $X_1(\omega_1)$, giving us eight unknowns and twelve equations. This makes the system overdetermined. In particular, simple matrix calculations reveal that, in general, we have no solution for X_0.

5. Stability under model perturbations

We now turn to the question of stability of the Föllmer–Schweizer decomposition. There are different kinds of perturbations one can consider, and for example, stability with respect to perturbations of V_N is considered in [Monat and Stricker 1995]. In this paper, we consider perturbations of the stock price process. For the stability analysis, as we work on finite probability spaces, the exact form of perturbations is not important, and we will suppose that there is a family of adapted stock price processes parametrized by ε, $(S^\varepsilon)_{\varepsilon\in(-\varepsilon_0,\varepsilon_0)}$ for some $\varepsilon_0 > 0$. An example of such a family corresponds to linear perturbations of the drift and volatility considered in the following section. Here we will only suppose that

$$S^\varepsilon \to S^0$$

in the sense that

$$\lim_{\varepsilon\to 0} S_n^\varepsilon(\omega) = S_n^0(\omega) \quad \text{for every } n \in \{0,\dots,N\} \text{ and } \omega \in \Omega. \tag{8}$$

The following result asserts that the Föllmer–Schweizer decomposition on finite probability spaces is stable under perturbations of the stock of the form (8).

Theorem 5.1. *On a finite probability space, let us consider a family of stock price processes* $((S_n^\varepsilon)_{n\in\{0,\dots,N\}})_{\varepsilon\in(-\varepsilon_0,\varepsilon_0)}$, *for some* $\varepsilon_0 > 0$, *satisfying* (8). *Let* V_N *be given. Then the corresponding family of the Föllmer–Schweizer decompositions*

$$V_N = V_0^\varepsilon + \sum_{j=1}^N \hat\vartheta_j^\varepsilon \Delta S_j^\varepsilon + L_N^\varepsilon, \quad \varepsilon \in (-\varepsilon_0, \varepsilon_0),$$

satisfies

$$\begin{aligned} &\lim_{\varepsilon\to 0} V_0^\varepsilon = V_0^0, \\ &\lim_{\varepsilon\to 0} L_n^\varepsilon = L_n^0, \quad n \in \{0,\dots,N\}, \\ &\lim_{\varepsilon\to 0} \hat\vartheta_n^\varepsilon = \hat\vartheta_n^0, \quad n \in \{1,\dots,N\}, \end{aligned} \tag{9}$$

where the equalities hold for every $\omega \in \Omega$. As a consequence, we also have

$$\lim_{\varepsilon \to 0} \sum_{j=1}^{n} \hat{\vartheta}_j^{\varepsilon} \Delta S_j^{\varepsilon} = \sum_{j=1}^{n} \hat{\vartheta}_j^{0} \Delta S_j^{0}, \quad n \in \{1, \dots, N\},\ \omega \in \Omega. \tag{10}$$

Proof. The proof goes recursively, backward in n. First, let us consider $n = N$. From (8), we get

$$\lim_{\varepsilon \to 0} \Delta S_N^{\varepsilon} = S_N^0, \quad \omega \in \Omega.$$

Further, working on a finite probability space, via the formal definition of conditional expectation, from (8), without any additional assumptions, we get

$$\lim_{\varepsilon \to 0} \mathbb{E}_{\mathcal{F}_{N-1}}[\Delta S_N^{\varepsilon}] = \mathbb{E}_{\mathcal{F}_{N-1}}[\Delta S_N^0].$$

As a consequence, we obtain

$$\begin{aligned}
\lim_{\varepsilon \to 0} \hat{\vartheta}_N^{\varepsilon} &= \lim_{\varepsilon \to 0} \frac{\mathrm{Cov}_{\mathcal{F}_{N-1}}[V_N, \Delta S_N^{\varepsilon}]}{\mathrm{Var}_{\mathcal{F}_{N-1}}[\Delta S_N^{\varepsilon}]} \\
&= \lim_{\varepsilon \to 0} \frac{\mathbb{E}_{\mathcal{F}_{N-1}}[(V_N - \mathbb{E}_{\mathcal{F}_{N-1}}[V_N])(\Delta S_N^{\varepsilon} - \mathbb{E}_{\mathcal{F}_{N-1}}[\Delta S_N^{\varepsilon}])]}{\mathbb{E}_{\mathcal{F}_{N-1}}[(\Delta S_N^{\varepsilon} - \mathbb{E}_{\mathcal{F}_{N-1}}[\Delta S_N^{\varepsilon}])^2]} \\
&= \frac{\mathbb{E}_{\mathcal{F}_{N-1}}[(V_N - \mathbb{E}_{\mathcal{F}_{N-1}}[V_N]) \lim_{\varepsilon \to 0}(\Delta S_N^{\varepsilon} - \mathbb{E}_{\mathcal{F}_{N-1}}[\Delta S_N^{\varepsilon}])]}{\mathbb{E}_{\mathcal{F}_{N-1}}[\lim_{\varepsilon \to 0}(\Delta S_N^{\varepsilon} - \mathbb{E}_{\mathcal{F}_{N-1}}[\Delta S_N^{\varepsilon}])^2]} \\
&= \frac{\mathbb{E}_{\mathcal{F}_{N-1}}[(V_N - \mathbb{E}_{\mathcal{F}_{N-1}}[V_N])(\lim_{\varepsilon \to 0} \Delta S_N^{\varepsilon} - \lim_{\varepsilon \to 0} \mathbb{E}_{\mathcal{F}_{N-1}}[\Delta S_N^{\varepsilon}])]}{\mathbb{E}_{\mathcal{F}_{N-1}}[(\lim_{\varepsilon \to 0} \Delta S_N^{\varepsilon} - \lim_{\varepsilon \to 0} \mathbb{E}_{\mathcal{F}_{N-1}}[\Delta S_N^{\varepsilon}])^2]} \\
&= \frac{\mathbb{E}_{\mathcal{F}_{N-1}}[(V_N - \mathbb{E}_{\mathcal{F}_{N-1}}[V_N])(\Delta S_N^{0} - \mathbb{E}_{\mathcal{F}_{N-1}}[\Delta S_N^{0}])]}{\mathbb{E}_{\mathcal{F}_{N-1}}[(\Delta S_N^{0} - \mathbb{E}_{\mathcal{F}_{N-1}}[\Delta S_N^{0}])^2]} \\
&= \frac{\mathrm{Cov}_{\mathcal{F}_{N-1}}[V_N, \Delta S_N^{0}]}{\mathrm{Var}_{\mathcal{F}_{N-1}}[\Delta S_N^{0}]} \\
&= \hat{\vartheta}_N^{0},
\end{aligned} \tag{11}$$

where the chain of equalities holds for every $\omega \in \Omega$. If $N = 1$, (11) implies the third equality in (9), and the remaining assertions of the theorem follow. If $N > 1$, defining

$$A_n^{\varepsilon} := V_N - \sum_{j=n+1}^{N} \hat{\vartheta}_j^{\varepsilon} \Delta S_j^{\varepsilon}, \quad n \in \{0, \dots, N-1\},\ \varepsilon \in (-\varepsilon_0, \varepsilon_0),$$

from (11), we get

$$\lim_{\varepsilon \to 0} A_{N-1}^{\varepsilon} = A_{N-1}^0, \quad \omega \in \Omega.$$

Consequently, similarly to (11), we obtain

$$\begin{aligned}\lim_{\varepsilon\to 0}\hat{\vartheta}^{\varepsilon}_{N-1} &= \lim_{\varepsilon\to 0}\frac{\mathrm{Cov}_{\mathcal{F}_{N-2}}[A^{\varepsilon}_{N-1},\,\Delta S^{\varepsilon}_{N-1}]}{\mathrm{Var}_{\mathcal{F}_{N-2}}[\Delta S^{\varepsilon}_{N-1}]}\\ &= \frac{\mathrm{Cov}_{\mathcal{F}_{N-2}}[\lim_{\varepsilon\to 0} A^{\varepsilon}_{N-1},\,\lim_{\varepsilon\to 0}\Delta S^{\varepsilon}_{N-1}]}{\mathrm{Var}_{\mathcal{F}_{N-2}}[\lim_{\varepsilon\to 0}\Delta S^{\varepsilon}_{N-1}]}\\ &= \frac{\mathrm{Cov}_{\mathcal{F}_{N-2}}[A^{0}_{N-1},\,\Delta S^{0}_{N-1}]}{\mathrm{Var}_{\mathcal{F}_{N-2}}[\Delta S^{0}_{N-1}]}\\ &= \hat{\vartheta}^{0}_{N-1},\quad \omega\in\Omega. \end{aligned} \tag{12}$$

Proceeding in such a manner, one can show that

$$\lim_{\varepsilon\to 0}\hat{\vartheta}^{\varepsilon}_{n} = \hat{\vartheta}^{0}_{n},\quad n\in\{1,\dots,N\},\ \omega\in\Omega,$$

which is the last equality in (9). In turn, this and (8) imply (10). Therefore, for every $\varepsilon\in(-\varepsilon_0,\varepsilon_0)$, by taking expectation in

$$V_N = V^{\varepsilon}_0 + \sum_{j=1}^{N}\hat{\vartheta}^{\varepsilon}_j\Delta S^{\varepsilon}_j + L^{\varepsilon}_N, \tag{13}$$

and using $\mathbb{E}[L^{\varepsilon}_N]=0$, $\varepsilon\in(-\varepsilon_0,\varepsilon_0)$, we deduce via (10) that $\lim_{\varepsilon\to 0}V^{\varepsilon}_0 = V^0_0$; i.e., the first equality in (9) holds. Consequently, as the left-hand side in (13) does not depend on ε, from (13), the convergence of V^{ε}_0 to V^0_0 and (10), we deduce that $\lim_{\varepsilon\to 0}L^{\varepsilon}_N = L^0_N$ for every $\omega\in\Omega$. Finally, as L^{ε}'s are $\mathbb{P}$-martingales, using $\mathbb{E}_{\mathcal{F}_n}[L^{\varepsilon}_N] = L^{\varepsilon}_n$, we conclude that

$$\lim_{\varepsilon\to 0}L^{\varepsilon}_n = L^0_n \quad \text{for every } n\in\{0,\dots,N\} \text{ and } \omega\in\Omega,$$

which is the second equality in (9). □

6. Asymptotic analysis

While stability tells us whether or not there is a convergence of the problem outputs under perturbations of the input data, the asymptotic analysis gives a quantitative estimate of how the problem responds to such perturbations. In order to make such estimates, assuming merely $S^{\varepsilon}\to S^0$ as in (8) from the previous section is not enough. We need to parametrize perturbations more precisely. Before stating our form of perturbations, one can also consider that, in practice, the dynamics of the stock price process is commonly decomposed into two parts. The first one is drift, or in discounted settings, it can also be stated as the market price of risk. This part is responsible for the trend of the stock. The second part captures the fluctuations, that is, how much can the stock fluctuate in a given interval. Therefore, one can

formulate the following dynamics of the stock for the base model:[2]

$$\Delta S_n^0 = \lambda_n + \sigma_n \Delta W_n = \lambda_n \Delta t + \sigma_n \Delta W_n, \quad n \in \{1, \dots, N\}, \tag{14}$$

where λ and σ are predictable processes, $\Delta t \equiv 1$ represents the change in time, and ΔW_n is an $\mathcal{F}_n$-measurable increment of a martingale with initial value 0, interpreted as an error or "noise" term, where we additionally suppose that the standard deviation of ΔW_n is $\Delta t = 1$ for normalization purposes. Equation (14) is also consistent with the so-called semimartingale decomposition of the stock price process, where a semimartingale can be defined to be a process that can be written as a sum of a martingale (noise term) and an adapted process (drift term). In the finite probability settings, in fact, the drift term in the semimartingale decomposition can be chosen to be predictable.

Remark 6.1. We observe that given any process S, and once the time step $\Delta t > 0$ is fixed (and is constant in this paper, for simplicity of notations), the processes λ and σ can be obtained as follows:

$$\begin{aligned} \lambda_n &= \frac{\mathbb{E}_{\mathcal{F}_{n-1}}[\Delta S_n]}{\Delta t}, \\ \sigma_n &= \sqrt{\mathrm{Var}_{\mathcal{F}_{n-1}}[\Delta S_n]}, \\ \Delta W_n &= \frac{\Delta S_n^0 - \lambda_n \Delta t}{\sigma_n} 1_{\{\sigma_n > 0\}}, \quad n \in \{1, \dots, N\}. \end{aligned}$$

For perturbations of the underlying dynamics in (14), we can consider simultaneous (or separate) distortions of both the drift and the noise terms in (14). Therefore, we now define model perturbations as

$$\Delta S_n^\varepsilon = (\lambda_n + \varepsilon \lambda_n') \Delta t + (\sigma_n + \varepsilon \sigma_n') \Delta W_n + \varepsilon \sigma_n'' \Delta W_n^\perp, \quad n \in \{1, \dots, N\},\ \varepsilon \in (-\varepsilon_0, \varepsilon_0),$$

where λ', σ', and σ'' are predictable processes, $\Delta W_n^\perp$ is an $\mathcal{F}_n$-measurable normalized noise term, which is conditionally uncorrelated from ΔW_n, that is,

$$\begin{aligned} \mathbb{E}_{\mathcal{F}_{n-1}}[\Delta W_n^\perp] &= 0, \\ \mathrm{Var}_{\mathcal{F}_{n-1}}[\Delta W_n^\perp] &= 1, \\ \mathrm{Cov}_{\mathcal{F}_{n-1}}[\Delta W_n, \Delta W_n^\perp] &= 0, \quad n \in \{1, \dots, N\}, \end{aligned}$$

and ε_0 is a strictly positive constant. With

$$\Delta S_n' := \lambda_n' \Delta t + \sigma_n' \Delta W_n + \sigma_n'' \Delta W_n^\perp, \quad n \in \{1, \dots, N\}, \tag{15}$$

we can rewrite the dynamics of the perturbed processes as

$$\Delta S_n^\varepsilon = \Delta S_n^0 + \varepsilon \Delta S_n', \quad n \in \{1, \dots, N\},\ \varepsilon \in (-\varepsilon_0, \varepsilon_0). \tag{16}$$

[2]Here the base model is the one that corresponds to $\varepsilon = 0$ in the notation of Sections 5 and 6.

Remark 6.2. Similarly to the process ΔS, if one starts from a given perturbation process $\Delta S'$, it can be represented in the form (15) as follows: In the notation of Remark 6.1, we have

$$\lambda'_n = \frac{\mathbb{E}_{\mathcal{F}_{n-1}}[\Delta S'_n]}{\Delta t}, \quad n \in \{1, \dots, N\},$$

and the computation of the remaining parameters goes along the lines of [Shreve 2004b, Example 2.3.3, p. 72]. Let us consider

$$\Delta S'_n - \mathbb{E}_{\mathcal{F}_{n-1}}[\Delta S'_n] = \sigma'_n \Delta W_n + (\Delta S'_n - \mathbb{E}_{\mathcal{F}_{n-1}}[\Delta S'_n] - \sigma'_n \Delta W_n), \quad n \in \{1, \dots, N\},$$

where σ'_n has to be determined in such a way that the term in the brackets is conditionally uncorrelated from ΔW_n. As $\mathrm{Var}_{\mathcal{F}_{n-1}}[\Delta W_n] = 1$, direct calculations give

$$\sigma'_n = \mathbb{E}_{\mathcal{F}_{n-1}}[\Delta W_n, \Delta S'_n - \mathbb{E}_{\mathcal{F}_{n-1}}[\Delta S'_n]] = \mathrm{Cov}_{\mathcal{F}_{n-1}}[\Delta W_n, \Delta S'_n], \quad n \in \{1, \dots, N\}.$$

Therefore, for σ''_n and $\Delta W^{\perp}_n$, we get

$$\sigma''_n = \sqrt{\mathrm{Var}_{\mathcal{F}_{n-1}}[\Delta S'_n - \mathbb{E}_{\mathcal{F}_{n-1}}[\Delta S'_n] - \sigma'_n \Delta W_n]} = \sqrt{\mathrm{Var}_{\mathcal{F}_{n-1}}[\Delta S'_n - \sigma'_n \Delta W_n]},$$

$$\Delta W^{\perp}_n = \frac{\Delta S'_n - \mathbb{E}_{\mathcal{F}_{n-1}}[\Delta S'_n] - \sigma'_n \Delta W_n}{\sigma''_n} 1_{\{\sigma''_n > 0\}}, \quad n \in \{1, \dots, N\}.$$

The following theorem gives the leading-order correction terms to the components of the Föllmer–Schweizer decomposition under perturbations of the form (16). Let us consider the associated family of the Föllmer–Schweizer decompositions

$$V_N = V_0^{\varepsilon} + \sum_{j=1}^{N} \hat{\vartheta}_j^{\varepsilon} \Delta S_j^{\varepsilon} + L_N^{\varepsilon}, \quad \varepsilon \in (-\varepsilon_0, \varepsilon_0), \tag{17}$$

and let us define recursively, backward in n, the process $\hat{\vartheta}'$, which will be proven in Theorem 6.3 to be the first-order correction to the optimal strategy, as

$$\hat{\vartheta}'_N := \frac{\mathrm{Cov}_{\mathcal{F}_{N-1}}[V_N, \Delta S'_N]}{\mathrm{Var}_{\mathcal{F}_{N-1}}[\Delta S_N^0]} - 2\frac{\mathrm{Cov}_{\mathcal{F}_{N-1}}[V_N, \Delta S_N^0]\,\mathrm{Cov}_{\mathcal{F}_{N-1}}[\Delta S_N^0, \Delta S'_N]}{(\mathrm{Var}_{\mathcal{F}_{N-1}}[\Delta S_N^0])^2},$$

$$\begin{aligned}\hat{\vartheta}'_n := \frac{1}{\mathrm{Var}_{\mathcal{F}_{n-1}}[\Delta S_n^0]} \Bigg(& \mathrm{Cov}_{\mathcal{F}_{n-1}}\Bigg[V_N - \sum_{j=n+1}^{N} \hat{\vartheta}_j^0 \Delta S_j^0, \Delta S'_n \Bigg] \\ & - \mathrm{Cov}_{\mathcal{F}_{n-1}}\Bigg[\sum_{j=n+1}^{N} \hat{\vartheta}'_j \Delta S_j^0 + \sum_{j=n+1}^{N} \hat{\vartheta}_j^0 \Delta S'_j, \Delta S_n^0 \Bigg] \Bigg) \\ & - 2\frac{\mathrm{Cov}_{\mathcal{F}_{n-1}}[V_N - \sum_{j=n+1}^{N} \hat{\vartheta}_j^0 \Delta S_j^0, \Delta S_n^0]\,\mathrm{Cov}_{\mathcal{F}_{N-1}}[\Delta S_n^0, \Delta S'_n]}{(\mathrm{Var}_{\mathcal{F}_{n-1}}[\Delta S_n^0])^2},\end{aligned} \tag{18}$$

where $n \in \{N-1, \dots, 1\}$.

Theorem 6.3. *On a finite probability space, let us consider a family of stock price processes* $((S_n^\varepsilon)_{n\in\{0,\dots,N\}})_{\varepsilon\in(-\varepsilon_0,\varepsilon_0)}$ *for some* $\varepsilon_0 > 0$, *where the increments,* $\Delta S_n^{\varepsilon\prime}$, *are given via* (16). *Let* V_N *be given. Then the components of the family of the Föllmer–Schweizer decompositions defined in* (17) *satisfy*

$$\begin{aligned}
\lim_{\varepsilon\to 0}\frac{V_0^\varepsilon - V_0^0}{\varepsilon} &= -\mathbb{E}\bigg[\sum_{j=1}^N \hat\vartheta_j^0 \Delta S_j' + \sum_{j=1}^N \hat\vartheta_j' \Delta S_j^0\bigg],\\
\lim_{\varepsilon\to 0}\frac{\hat\vartheta_n^\varepsilon - \hat\vartheta_n^0}{\varepsilon} &= \hat\vartheta_n', \quad n\in\{1,\dots,N\},\ \omega\in\Omega,\\
\lim_{\varepsilon\to 0}\frac{L_n^\varepsilon - L_n^0}{\varepsilon} &\\
= -\mathbb{E}_{\mathcal{F}_n}\bigg[\sum_{j=1}^N \hat\vartheta_j' \Delta S_j^0 - \mathbb{E}\bigg[\sum_{j=1}^N \hat\vartheta_j' \Delta S_j^0\bigg]\bigg] &- \mathbb{E}_{\mathcal{F}_n}\bigg[\sum_{j=1}^N \hat\vartheta_j^0 \Delta S_j' - \mathbb{E}\bigg[\sum_{j=1}^N \hat\vartheta_j^0 \Delta S_j'\bigg]\bigg],\\
n\in\{0,\dots,N\},\ \omega\in\Omega,
\end{aligned} \tag{19}$$

where ϑ' *is defined in* (18). *We also have*

$$\begin{aligned}
\lim_{\varepsilon\to 0}\frac{\sum_{j=1}^n \hat\vartheta_j^\varepsilon \Delta S_j^\varepsilon - \sum_{j=1}^n \hat\vartheta_j^0 \Delta S_j^0}{\varepsilon}&\\
= \sum_{j=1}^n \hat\vartheta_j^0 \Delta S_j' + \sum_{j=1}^n \hat\vartheta_j' \Delta S_j^0, &\quad n\in\{1,\dots,N\},\ \omega\in\Omega.
\end{aligned} \tag{20}$$

Proof. We will investigate

$$\lim_{\varepsilon\to 0}\frac{\hat\vartheta_n^\varepsilon - \hat\vartheta_n^0}{\varepsilon}$$

first. For $n = N$, we have

$$\lim_{\varepsilon\to 0}\frac{\hat\vartheta_n^\varepsilon - \hat\vartheta_n^0}{\varepsilon} = \lim_{\varepsilon\to 0}\bigg(\frac{\operatorname{Cov}_{\mathcal{F}_{N-1}}[V_N, \Delta S_N^\varepsilon]}{\operatorname{Var}_{\mathcal{F}_{N-1}}[\Delta S_N^\varepsilon]} - \frac{\operatorname{Cov}_{\mathcal{F}_{N-1}}[V_N, \Delta S_N^\varepsilon]}{\operatorname{Var}_{\mathcal{F}_{N-1}}[\Delta S_N^\varepsilon]}\bigg)\frac{1}{\varepsilon}. \tag{21}$$

Using the definition of conditional expectation, and Theorem 5.1,[3] we get

$$\lim_{\varepsilon\to 0}\frac{\operatorname{Cov}_{\mathcal{F}_{N-1}}[V_n, \Delta S_N^\varepsilon] - \operatorname{Cov}_{\mathcal{F}_{N-1}}[V_n, \Delta S_N^0]}{\varepsilon} = \operatorname{Cov}_{\mathcal{F}_{N-1}}[V_N, \Delta S_N'].$$

Similarly, we deduce that

$$\lim_{\varepsilon\to 0}\frac{\operatorname{Var}_{\mathcal{F}_{N-1}}[\Delta S_N^\varepsilon] - \operatorname{Var}_{\mathcal{F}_{N-1}}[\Delta S_N^0]}{\varepsilon} = 2\operatorname{Cov}_{\mathcal{F}_{N-1}}[\Delta S_N^0, \Delta S_N'].$$

[3]Below, we apply the assertions of Theorem 5.1 at a family of points ε near the origin; this however causes no difficulty by relabeling S^ε's.

Therefore, in (21), we obtain

$$\lim_{\varepsilon\to 0}\frac{\hat\vartheta_n^\varepsilon-\hat\vartheta_n^0}{\varepsilon}$$
$$=\frac{\mathrm{Cov}_{\mathcal{F}_{N-1}}[V_N,\Delta S'_N]\,\mathrm{Var}_{\mathcal{F}_{N-1}}[\Delta S_N^0]-2\,\mathrm{Cov}_{\mathcal{F}_{N-1}}[V_N,\Delta S_N^0]\,\mathrm{Cov}_{\mathcal{F}_{N-1}}[\Delta S_N^0,\Delta S'_N]}{(\mathrm{Var}_{\mathcal{F}_{N-1}}[\Delta S_N^0])^2},$$

which is precisely $\hat\vartheta'_N$. If $N=1$, this completes the proof for

$$\lim_{\varepsilon\to 0}\frac{\hat\vartheta_n^\varepsilon-\hat\vartheta_n^0}{\varepsilon}=\vartheta'_n$$

for every n. If $N>1$, defining

$$A_n^\varepsilon:=V_N-\sum_{j=n+1}^N\hat\vartheta_j^\varepsilon\Delta S_j^\varepsilon,\quad n\in\{0,\dots,N-1\},\ \varepsilon\in(-\varepsilon_0,\varepsilon_0),$$

and using Theorem 5.1, we get recursively, backward in n, the following chain of equalities. First, for $n=N-1$, we obtain

$$\lim_{\varepsilon\to 0}\frac{A_n^\varepsilon-A_n^0}{\varepsilon}=-\sum_{j=n+1}^N\hat\vartheta'_j\Delta S_j^0-\sum_{j=n+1}^N\hat\vartheta_j^0\Delta S'_j.$$

Therefore, using Theorem 5.1 again, we obtain

$$\lim_{\varepsilon\to 0}\frac{\mathrm{Cov}_{\mathcal{F}_{n-1}}[A_n^\varepsilon,\Delta S_n^\varepsilon]-\mathrm{Cov}_{\mathcal{F}_{n-1}}[A_n^0,\Delta S_n^0]}{\varepsilon}$$
$$=\mathrm{Cov}_{\mathcal{F}_{n-1}}\Big[-\sum_{j=n+1}^N\hat\vartheta'_j\Delta S_j^0-\sum_{j=n+1}^N\hat\vartheta_j^0\Delta S'_j,\Delta S_n^0\Big]+\mathrm{Cov}_{\mathcal{F}_{n-1}}[A_n^0,\Delta S'_n]$$
$$=\mathrm{Cov}_{\mathcal{F}_{n-1}}\Big[-\sum_{j=n+1}^N\hat\vartheta'_j\Delta S_j^0-\sum_{j=n+1}^N\hat\vartheta_j^0\Delta S'_j,\Delta S_n^0\Big]+\mathrm{Cov}_{\mathcal{F}_{n-1}}\Big[V_N-\sum_{j=n+1}^N\hat\vartheta_j^0\Delta S_j^0,\Delta S'_n\Big],$$

and thus, in (21), we conclude that

$$\lim_{\varepsilon\to 0}\frac{\hat\vartheta_n^\varepsilon-\hat\vartheta_n^0}{\varepsilon}=\frac{1}{\mathrm{Var}_{\mathcal{F}_{n-1}}[\Delta S_n^0]}\Big(\mathrm{Cov}_{\mathcal{F}_{n-1}}\Big[V_N-\sum_{j=n+1}^N\hat\vartheta_j^0\Delta S_j^0,\Delta S'_n\Big]$$
$$-\mathrm{Cov}_{\mathcal{F}_{n-1}}\Big[\sum_{j=n+1}^N\hat\vartheta'_j\Delta S_j^0+\sum_{j=n+1}^N\hat\vartheta_j^0\Delta S'_j,\Delta S_n^0\Big]\Big)$$
$$-\frac{2\,\mathrm{Cov}_{\mathcal{F}_{n-1}}\big[V_N-\sum_{j=n+1}^N\hat\vartheta_j^0\Delta S_j^0,\Delta S_n^0\big]\,\mathrm{Cov}_{\mathcal{F}_{N-1}}[\Delta S_n^0,\Delta S'_n]}{(\mathrm{Var}_{\mathcal{F}_{n-1}}[\Delta S_n^0])^2},$$

which is exactly $\hat{\vartheta}'_n$ from (18) for $n = N-1$. Proceeding this way, we can establish (18) for every $n \geq 1$. This completes the proof of the third equality in (19). Now, (20) follows from the third equality in (19) and Theorem 5.1.

To establish the first two equalities in (19), we proceed as follows. For every $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$, taking the expectation in (17) and observing that the left-hand side does not depend on ε, as well as that $\mathbb{E}[L_N^\varepsilon] = 0$, we get

$$V_0^\varepsilon + \mathbb{E}\bigg[\sum_{j=1}^N \hat{\vartheta}_j^\varepsilon \Delta S_j^\varepsilon\bigg] = V_0^0 + \mathbb{E}\bigg[\sum_{j=1}^N \hat{\vartheta}_j^0 \Delta S_j^0\bigg], \quad \varepsilon \in (-\varepsilon_0, \varepsilon_0).$$

Collecting the terms and dividing by $\varepsilon \neq 0$, we obtain

$$\frac{V_0^\varepsilon - V_0^0}{\varepsilon} = \frac{\mathbb{E}\big[\sum_{j=1}^N \hat{\vartheta}_j^0 \Delta S_j^0 - \sum_{j=1}^N \hat{\vartheta}_j^\varepsilon \Delta S_j^\varepsilon\big]}{\varepsilon}, \quad \varepsilon \in (-\varepsilon_0, 0) \cup (0, \varepsilon_0).$$

Taking the limit as $\varepsilon \to 0$, and using (20), we conclude that

$$\lim_{\varepsilon \to 0} \frac{V_0^\varepsilon - V_0^0}{\varepsilon} = -\mathbb{E}\bigg[\sum_{j=1}^N \hat{\vartheta}_j^0 \Delta S'_j + \sum_{j=1}^N \hat{\vartheta}'_j \Delta S_j^0\bigg],$$

which is precisely the first equality in (19).

To obtain the remaining assertion in (19), we observe that from (17), (20) and the other two assertions in (19), we immediately obtain

$$\begin{aligned}
\lim_{\varepsilon \to 0} \frac{L_N^\varepsilon - L_N^0}{\varepsilon} &= -\lim_{\varepsilon \to 0} \frac{V_0^\varepsilon - V_0^0}{\varepsilon} - \lim_{\varepsilon \to 0} \frac{\sum_{j=1}^N \hat{\vartheta}_j^\varepsilon \Delta S_j^\varepsilon - \sum_{j=1}^N \hat{\vartheta}_j^0 \Delta S_j^0}{\varepsilon} \\
&= \mathbb{E}\bigg[\sum_{j=1}^N \hat{\vartheta}_j^0 \Delta S'_j + \sum_{j=1}^N \hat{\vartheta}'_j \Delta S_j^0\bigg] - \sum_{j=1}^N \hat{\vartheta}_j^0 \Delta S'_j - \sum_{j=1}^N \hat{\vartheta}'_j \Delta S_j^0 \\
&= -\bigg(\sum_{j=1}^N \hat{\vartheta}_j^0 \Delta S'_j - \mathbb{E}\bigg[\sum_{j=1}^N \hat{\vartheta}_j^0 \Delta S'_j\bigg]\bigg) - \bigg(\sum_{j=1}^N \hat{\vartheta}'_j \Delta S_j^0 - \mathbb{E}\bigg[\sum_{j=1}^N \hat{\vartheta}'_j \Delta S_j^0\bigg]\bigg).
\end{aligned}$$

If $N = 1$, this completes the proof. If $N > 1$, for $n < N$, we have $\mathbb{E}_{\mathcal{F}_n}[L_N^\varepsilon] = L_n^\varepsilon$, $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$. Therefore, using (17) and taking the conditional expectation, we obtain

$$\begin{aligned}
\frac{L_n^\varepsilon - L_n^0}{\varepsilon} &= \frac{\mathbb{E}_{\mathcal{F}_n}[L_N^\varepsilon - L_N^0]}{\varepsilon} \\
&= -\frac{V_0^\varepsilon - V_0^0}{\varepsilon} - \frac{\mathbb{E}_{\mathcal{F}_n}\big[\sum_{j=1}^N \hat{\vartheta}_j^\varepsilon \Delta S_j^\varepsilon - \sum_{j=1}^N \hat{\vartheta}_j^0 \Delta S_j^0\big]}{\varepsilon}.
\end{aligned}$$

Taking the limit, and using (20) and the first equality in (19), we conclude that

$$\lim_{\varepsilon\to 0}\frac{L_n^\varepsilon - L_n^0}{\varepsilon} = -\mathbb{E}_{\mathcal{F}_n}\left[\sum_{j=1}^N \hat{\vartheta}_j'\Delta S_j^0 - \mathbb{E}\left[\sum_{j=1}^N \hat{\vartheta}_j'\Delta S_j^0\right]\right] - \mathbb{E}_{\mathcal{F}_n}\left[\sum_{j=1}^N \hat{\vartheta}_j^0\Delta S_j' - \mathbb{E}\left[\sum_{j=1}^N \hat{\vartheta}_j^0\Delta S_j'\right]\right].$$

This completes the proof of the theorem. □

References

[Brigo and Mercurio 2006] D. Brigo and F. Mercurio, *Interest rate models — theory and practice*, 2nd ed., Springer, 2006. MR Zbl

[Föllmer and Schweizer 1989] H. Föllmer and M. Schweizer, "Hedging by sequential regression: an introduction to the mathematics of option trading", *ASTIN Bull.* **18**:2 (1989), 147–160.

[Kramkov and Sîrbu 2006a] D. Kramkov and M. Sîrbu, "On the two-times differentiability of the value functions in the problem of optimal investment in incomplete markets", *Ann. Appl. Probab.* **16**:3 (2006), 1352–1384. MR Zbl

[Kramkov and Sîrbu 2006b] D. Kramkov and M. Sîrbu, "Sensitivity analysis of utility-based prices and risk-tolerance wealth processes", *Ann. Appl. Probab.* **16**:4 (2006), 2140–2194. MR Zbl

[Monat and Stricker 1995] P. Monat and C. Stricker, "Föllmer–Schweizer decomposition and mean-variance hedging for general claims", *Ann. Probab.* **23**:2 (1995), 605–628. MR Zbl

[Mostovyi 2020] O. Mostovyi, "Asymptotic analysis of the expected utility maximization problem with respect to perturbations of the numéraire", *Stochastic Process. Appl.* **130**:7 (2020), 4444–4469. MR Zbl

[Mostovyi and Sîrbu 2019] O. Mostovyi and M. Sîrbu, "Sensitivity analysis of the utility maximisation problem with respect to model perturbations", *Finance Stoch.* **23**:3 (2019), 595–640. MR Zbl

[Schweizer 1994] M. Schweizer, "Approximating random variables by stochastic integrals", *Ann. Probab.* **22**:3 (1994), 1536–1575. MR Zbl

[Schweizer 1995] M. Schweizer, "Variance-optimal hedging in discrete time", *Math. Oper. Res.* **20**:1 (1995), 1–32. MR Zbl

[Shreve 2004a] S. E. Shreve, *Stochastic calculus for finance, I*: *The binomial asset pricing model*, Springer, 2004. MR Zbl

[Shreve 2004b] S. E. Shreve, *Stochastic calculus for finance, II*: *Continuous-time models*, Springer, 2004. MR Zbl

Received: 2020-01-11 Accepted: 2020-05-23

sboese@vassar.edu *Vassar College, Poughkeepsie, NY, United States*

txxcui@gmail.com *Carnegie Mellon University, Pittsburgh, PA, United States*

sdjohnston@willamette.edu *Willamette University, Salelm, OR, United States*

sylvie.vega-molino@uib.no *Department of Mathematics, University of Connecticut, Storrs, CT, United States*

oleksii.mostovyi@uconn.edu *Department of Mathematics, University of Connecticut, Storrs, CT, United States*

INVOLVE 13:4 (2020)
https://doi.org/10.2140/involve.2020.13.625

Eigenvalues of the sum and product of anticommuting matrices

Vadim Ponomarenko and Louis Selstad

(Communicated by Chi-Kwong Li)

Improving on a recent result of Zhong, we characterize the eigenvalues of AB and $A+B$ for square matrices A, B satisfying $AB+BA=0$.

Square matrices A, B are called anticommuting if $AB=-BA$. Such matrices are important in mathematical physics, e.g., as Pauli spin matrices. They are of continued mathematical interest; see, e.g., [Hrubeš 2016; Shapiro and Martin 1998; Taussky 1970]. The recent paper [Zhong 2017] presented the following theorem.

Theorem 1 (Zhong). *Let A, B be square matrices with $\sigma(A)=\{\lambda_1,\dots,\lambda_n\}$, $\sigma(B)=\{\mu_1,\dots,\mu_n\}$. Suppose that A, B anticommute, and for each eigenvalue λ_i of A the algebraic multiplicity equals the geometric multiplicity. Then*

(1) $\sigma(AB)\subseteq\{\lambda_j\mu_k i : \text{all } j,k\}$, *and*

(2) $\sigma(A+B)\subseteq\{\pm\sqrt{\lambda_j^2+\mu_k^2} : \text{all } j,k\}$.

We improve on the previous theorem by slightly weakening the hypotheses, and finding $\sigma(AB)$ and $\sigma(A+B)$ exactly (in addition to various structural results). We will put A into Jordan canonical form, which will impose a (to be defined) correspondence between eigenvalues of A and of B. Our main result is:

Main Theorem. *Let A, B be square matrices with $\sigma(A)=\{\lambda_1,\dots,\lambda_n\}$, $\sigma(B)=\{\mu_1,\dots,\mu_n\}$. Suppose that A, B anticommute, and for each nonzero eigenvalue λ_i of A the algebraic multiplicity equals the geometric multiplicity. Then*

(1) $\sigma(AB)=\{\lambda_j\mu_k i : \text{corresponding } j,k\}$, *and*

(2) $\sigma(A+B)=\{\pm\sqrt{\lambda_j^2+\mu_k^2} : \text{corresponding } j,k\}$.

Proof. See Corollaries 7 and 9, to follow. □

MSC2010: 15A18.
Keywords: anticommutative, eigenvalues, matrix spectrum.

Before proceeding, we set some notation. We define $\mathbb{C}^\bullet$ as $\mathbb{C}\setminus\{0\}$. Given a square $n\times n$ matrix M, we recall the characteristic polynomial $p_M(t)=|tI_n-M|$. A key property of characteristic polynomials is that $p_{AB}(t)=p_{BA}(t)$. If $AB+BA=0$, then

$$\begin{aligned}p_{AB}(t)&=p_{BA}(t)=p_{-AB}(t)\\&=|tI_n+AB|=(-1)^n|(-t)I_n-AB|=(-1)^n p_{AB}(-t).\end{aligned}$$

Consequently, for even (resp. odd) n, the function $p_{AB}(t)$ is even (resp. odd). Since $p_{AB}(t)$ is a polynomial, it must therefore be a sum of monomials in only even (resp. odd) powers of t. In general, $p_{AB}(t)$ cannot be determined from $p_A(t)$ and $p_B(t)$.

We now present a familiar definition, generalized to nonsquare matrices.

Definition 2. Let $m,n\in\mathbb{N}$, and let C be an $m\times n$ matrix. We say that C is *upper triangular* if it satisfies $C_{i,j}=0$ for all i, j with $j-i<\max(0,n-m)$.

For an upper triangular matrix C, its nonzero entries are confined to a triangle in the upper-right corner whose sides are horizontal, vertical, and of slope -1. This triangle has one corner at $C_{1,n}$ and the other corners at either $\{C_{1,1},C_{n,n}\}$ (if $m\geq n$) or at $\{C_{1,n-m+1},C_{m,n}\}$ (if $m\leq n$). For $m=n$, this coincides with the usual definition of "upper triangular".

The product of upper triangular matrices satisfies a useful property.

Theorem 3. *Let $m,n\in\mathbb{N}$. Let F be an upper triangular $m\times n$ matrix, and E an upper triangular $n\times m$ matrix. Then FE is an $m\times m$ matrix satisfying $(FE)_{i,j}=0$ for all i, j with $j-i<|n-m|$. In particular, if $n\neq m$, then FE is strictly upper triangular, and hence satisfies $p_{FE}(t)=t^m$.*

Proof. Note that $(FE)_{i,j}=\sum_{k=1}^n F_{i,k}E_{k,j}$. Suppose first that $n\geq m$, and $j-i<n-m$. Then $F_{i,k}=0$ for all i,k with $k-i<n-m$. Also, $E_{k,j}=0$ for all k, j with $j-k<0$. Hence, if $k>j$, then $E_{k,j}=0$; if instead $k\leq j$, then $k-i\leq j-i<n-m$, so $F_{i,k}=0$. Thus $(FE)_{i,j}=0$. Suppose now that $n<m$, and $j-i<m-n$. Then $F_{i,k}=0$ for all i,k with $k-i<0$. Also, $E_{k,j}=0$ for all k, j with $j-k<m-n$. Hence, if $k<i$, then $F_{i,k}=0$; if instead $k\geq i$, then $j-k\leq j-i<m-n$, so $E_{k,j}=0$. Thus $(FE)_{i,j}=0$. □

We recall the (square) upper shift matrix U_n (of size $n\in\mathbb{N}$), which has ones on the superdiagonal and zeroes elsewhere. Using the Kronecker delta function, we may define $(U_n)_{i,j}=\delta_{i+1,j}$. For future use, we define matrix $U_n^\star$ to be any integer matrix whose entries satisfy $0\leq(U_n^\star)_{i,j}\leq(U_n)_{i,j}$. We next recall (square) Jordan blocks, which exist for any eigenvalue α and any size $n\in\mathbb{N}$, defined as $J_n(\alpha)=\alpha I_n+U_n$.

We recall that a matrix is in Jordan canonical form if it is a block diagonal matrix, whose diagonal blocks are each Jordan blocks (of any size and any eigenvalues).

The celebrated Jordan canonical form theorem states that every square matrix is similar to one in Jordan form. We now apply this similarity to the anticommutative relation. If $AB + BA = 0$, choose P so that PAP^{-1} is in Jordan form. We have

$$PAP^{-1}PBP^{-1} + PBP^{-1}PAP^{-1} = P0P^{-1} = 0.$$

Set $A' = PAP^{-1}$ and $B' = PBP^{-1}$; we have $A'B' + B'A' = 0$. A is similar to A' and B is similar to B', and similarity preserves characteristic polynomials (and hence eigenvalues). Henceforth we simply assume that $AB + BA = 0$ and that A is in Jordan form.

We now partition A, B into blocks, based on the block diagonal structure of the Jordan form of A. Considering an arbitrary $m \times n$ block C of B we find that the anticommuting relation forces $J_m(\alpha)C + CJ_n(\beta) = 0$ for some diagonal Jordan blocks $J_m(\alpha)$, $J_n(\beta)$ of A. This forces most such blocks C to be zero, and imposes upper triangularity on the rest.

Theorem 4. *Let $m, n \in \mathbb{N}$, and let C be an $m \times n$ matrix. Suppose that*

$$J_m(\alpha)C + CJ_n(\beta) = 0$$

for some $\alpha, \beta \in \mathbb{C}$. If $\alpha + \beta \neq 0$, then $C = 0$. If instead $\alpha + \beta = 0$, then C must be upper triangular.

Proof. Note that $(U_m C)_{m,j} = 0$, and $(U_m C)_{i,j} = C_{i+1,j}$ for $i < m$. Similarly, $(CU_n)_{i,1} = 0$, and $(CU_n)_{i,j} = C_{i,j-1}$ otherwise. Conventionally, define $C_{i,j} = 0$ if $i > m$ or $j < 1$. We have

$$0 = J_m(\alpha)C + CJ_n(\beta) = (\alpha + \beta)C + U_m C + CU_n.$$

Now, $U_m C$ moves the rows of C upward, inserting a zero row, while CU_nmoves the columns of C to the right, inserting a zero column. Hence, in $U_m C + CU_n$, the southwest frontier of C must move northeast.

We first consider $\alpha + \beta \neq 0$. Set S to be the set of those $(i, j) \in \mathbb{N}^2 \cap [1, m] \times [1, n]$ which correspond to a nonzero entry in C. Suppose, by way of contradiction, that S is nonempty. Take $(i, j) \in K'$ that is minimal with respect to $j - i$, i.e., on the southwest frontier of C. If there is more than one with this minimal $j - i$, take the one with maximal i. Look at the (i, j)-entry of $0 = (\alpha + \beta)C + U_m C + CU_n$. We have

$$0 = (\alpha + \beta)C_{i,j} + C_{i+1,j} + C_{i,j-1}.$$

Either since $i + 1 > m$ or since $j - (i + 1) < j - i$, we must have $C_{i+1,j} = 0$. Either since $j - 1 < 1$ or since $(j - 1) - i < j - i$, we must have $C_{i,j-1} = 0$. In other words, the southwest frontier of $U_m C + CU_n$ has moved away from (i, j). Hence $0 = (\alpha + \beta)C_{i,j}$. But since $\alpha + \beta \neq 0$, we have $C_{i,j} = 0$, which is a contradiction. Consequently S is empty.

We now consider the case of $\alpha+\beta=0$. First we take the case $m \geq n$. Set K to be the set of those $(i, j) \in \mathbb{N}^2 \cap [1, m] \times [1, n]$ such that $j - i < 0$. Let K' be those elements of K which correspond to a nonzero entry in C. Suppose, by way of contradiction, that K' is nonempty. Take $(i, j) \in K'$ that is minimal with respect to $j-i$. If there is more than one with this minimal $j-i$, take the one with minimal i. We have

$$0 = (\alpha+\beta)C + U_m C + C U_n = U_m C + C U_n.$$

Note that if $i = 1$ then $j < 1$, which is impossible. Hence $i \geq 2$. Look at the $(i-1, j)$-entry of this matrix. We have $0 = C_{i,j} + C_{i-1,j-1}$. Either since $j < 1$, or since $(j-1)-(i-1) = j-i$ and $i-1 < i$, we must have $C_{i-1,j-1} = 0$. But then $C_{i,j} = 0$, a contradiction. Hence K' is empty.

Lastly, we consider the case $m > n$. Set K to be the set of those $(i, j) \in \mathbb{N}^2 \cap [1, m] \times [1, n]$ such that $j - i < n - m$. Let K' be those elements of K which correspond to a nonzero entry in C. Suppose, by way of contradiction, that K' is nonempty. Take $(i, j) \in K'$ that is minimal with respect to $j-i$. If there is more than one with this minimal $j-i$, take the one with maximal i. We have

$$0 = (\alpha+\beta)C + U_m C + C U_n = U_m C + C U_n.$$

Note that if $j = n$ then $-i < -m$, or $i > m$, which is impossible. Hence $j \leq n-1$. Look at the $(i, j+1)$-entry of this matrix. We have $0 = C_{i+1,j+1} + C_{i,j}$. Either since $i+1 > m$, or since $(j+1)-(i+1) = j-i$ and $i+1 > i$, we must have $C_{i+1,j+1} = 0$. But then $C_{i,j} = 0$, a contradiction. Hence K' is empty. □

It turns out that the upper triangular matrices C from Theorem 4 have additional banded structure, which we will not explore.

We now impose a particular order to the diagonal Jordan blocks of A. By reordering rows and columns if necessary, we collect together all of the Jordan blocks of eigenvalue 0 (if any) into one big block $A(0) = U_n^\star$ for some $n \in \mathbb{N}_0$ (here n is the algebraic multiplicity of eigenvalue 0 in A). The corresponding diagonal big block of B, which we call $B(0)$, is upper triangular by Theorem 4.

For each nonzero eigenvalue α, we also collect together all of the Jordan blocks of $\pm\alpha$ into one big block $A(\alpha)$, which has n copies of α on the diagonal, followed by m copies of $-\alpha$. By reversing the roles of α, $-\alpha$ if necessary, we assume that $n \geq m$. We have

$$A(\alpha) = \begin{pmatrix} \alpha I_n + U_n^\star & 0 \\ 0 & -\alpha I_m + U_m^\star \end{pmatrix}.$$

We now repartition A, B into big blocks, based on the block diagonal structure induced by these big blocks. We denote by

$$B(\alpha) = \begin{pmatrix} 0 & E \\ F & 0 \end{pmatrix}$$

the diagonal big block of B corresponding to $A(\alpha)$. We call this the big block Jordan form for B. By Theorem 4 again, B will now also be block diagonal. Hence $p_B(t)$ is just the product of all the $p_{B(\alpha)}(t)$.

These big blocks induce a correspondence between the eigenvalues of $A(\alpha)$, namely $\pm\alpha$, and the eigenvalues of $B(\alpha)$.

We can now compute the characteristic polynomial of $B(\alpha)$ (for nonzero α).

Theorem 5. *Let $m, n \in \mathbb{N}$ with $n \geq m$. Suppose E is an $n \times m$ matrix and F is an $m \times n$ matrix. Set*

$$B(\alpha) = \begin{pmatrix} 0 & E \\ F & 0 \end{pmatrix}.$$

Then $p_{B(\alpha)}(t) = t^{n-m} p_{FE}(t^2)$.

Proof. We calculate

$$p_{B(\alpha)}(t) = |tI - B(\alpha)| = \begin{vmatrix} tI_n & -E \\ -F & tI_m \end{vmatrix}.$$

Since tI_n is invertible, this equals

$$\begin{aligned} |tI_n|\,|tI_m - (-F)(tI_n)^{-1}(-E)| &= t^n |tI_m - t^{-1}FE| \\ &= t^n |t^{-1}I_m|\,|t^2 I_m - FE| = t^{n-m} p_{FE}(t^2). \quad \square \end{aligned}$$

If $n > m$, then by Theorem 3, $p_{FE}(t) = t^m$. Hence, $p_{B(\alpha)} = t^{n-m}t^{2m} = t^{n+m}$, and $B(\alpha)$ has only 0 as an eigenvalue. If we assume instead that big block $B(\alpha)$ has a nonzero eigenvalue, then $n = m$. This makes $p_{A(\alpha)}(t) = (t-\alpha)^m(t+\alpha)^m = (t^2-\alpha^2)^m$ and $p_{B(\alpha)}(t) = p_{FE}(t^2)$ both even. If we assume that A is invertible and that every big block $B(\alpha)$ has a nonzero eigenvalue, then $p_A(t)$ and $p_B(t)$ must both be even.

We now focus our attention on the case when big block $A(\alpha)$ is diagonalizable for nonzero α. In this case, the geometric multiplicity and algebraic multiplicity of α coincide.

Theorem 6. *Let $\alpha \in \mathbb{C}^\bullet$, and let $m, n \in \mathbb{N}$. Suppose we have*

$$A(\alpha) = \begin{pmatrix} \alpha I_n & 0 \\ 0 & -\alpha I_m \end{pmatrix}, \quad B(\alpha) = \begin{pmatrix} 0 & E \\ F & 0 \end{pmatrix}.$$

Then

$$\begin{aligned} p_{A(\alpha)B(\alpha)}(t) &= (i\alpha)^{n+m} p_{B(\alpha)}\left(\frac{t}{i\alpha}\right), \\ p_{A(\alpha)+B(\alpha)}(t) &= (t-\alpha)^{n-m} p_{FE}(t^2 - \alpha^2). \end{aligned}$$

Proof. We calculate

$$p_{A(\alpha)B(\alpha)}(t) = |tI - A(\alpha)B(\alpha)| = \begin{vmatrix} tI_n & -\alpha E \\ \alpha F & tI_m \end{vmatrix}.$$

Since tI_n is invertible, this equals

$$\begin{aligned}|tI_n||tI_m - (\alpha F)(tI_n)^{-1}(-\alpha E)| &= t^n|tI_m - (-\alpha^2 t^{-1})FE| \\ &= t^n|(-\alpha^2)t^{-1}I_m|\left|\left(-\frac{t^2}{\alpha^2}I_m - FE\right)\right| = t^{n-m}(-\alpha^2)^m p_{FE}\left(\left(\frac{t}{i\alpha}\right)^2\right) \\ &= \left(\frac{t}{i\alpha}\right)^{n-m}(i\alpha)^{n-m}(i\alpha)^{2m}p_{FE}\left(\left(\frac{t}{i\alpha}\right)^2\right) = (i\alpha)^{n+m}p_{B(\alpha)}\left(\frac{t}{i\alpha}\right).\end{aligned}$$

We now calculate

$$p_{A(\alpha)+B(\alpha)}(t) = |tI - A(\alpha) - B(\alpha)| = \begin{vmatrix}(t-\alpha)I_n & -E \\ F & (t+\alpha)I_m\end{vmatrix}.$$

Since $(t-\alpha)I_n$is invertible, this equals

$$\begin{aligned}|(t-\alpha)I_n||(t+\alpha)I_m - F((t-\alpha)I_n)^{-1}E| &= (t-\alpha)^n|(t+\alpha)I_m - (t-\alpha)^{-1}FE| \\ &= (t-\alpha)^n|(t-\alpha)^{-1}I_m||(t+\alpha)(t-\alpha)I_m - FE| \\ &= (t-\alpha)^{n-m}p_{FE}(t^2-\alpha^2).\end{aligned}$$

□

From our characteristic polynomial calculations, we can derive the eigenvalues of $A(\alpha)B(\alpha)$ and $A(\alpha)+B(\alpha)$ from the eigenvalues of $A(\alpha)$ and $B(\alpha)$.

Corollary 7. *Let $\alpha \in \mathbb{C}^\bullet$, and let $m, n \in \mathbb{N}$. Suppose we have*

$$A(\alpha) = \begin{pmatrix}\alpha I_n & 0 \\ 0 & -\alpha I_m\end{pmatrix}, \quad B(\alpha) = \begin{pmatrix}0 & E \\ F & 0\end{pmatrix}.$$

Let $\{\lambda_k\}$ denote the eigenvalues of $B(\alpha)$, with multiplicity. Then the eigenvalues of $A(\alpha)B(\alpha)$ are exactly $\{\pm i\alpha\lambda_k\}$, and the eigenvalues of $A(\alpha)+B(\alpha)$ are exactly $\{\pm\sqrt{\alpha^2+\lambda_k^2}\}$.

Proof. Let

$$p_{FE}(t) = \prod_{k=1}^{m}(t-\mu_k),$$

where the $\mu_k$are the eigenvalues of FE, not assumed distinct. Then

$$p_{B(\alpha)}(t) = t^{n-m}\prod_{k=1}^{m}(t^2-\mu_k);$$

hence the eigenvalues of $B(\alpha)$ consist of $n-m$ copies of 0, and $\pm\sqrt{\mu_k}$ for each eigenvalue of FE. We can interpret the complex square root as a principal value, but it doesn't matter since we get both square roots.

Now

$$p_{A(\alpha)B(\alpha)}(t) = \left(\frac{t}{i\alpha}\right)^{n-m} (i\alpha)^{n+m} \prod_{k=1}^{m} \left(\left(\frac{t}{i\alpha}\right)^2 - \mu_k\right) = t^{n-m} \prod_{k=1}^{m} (t^2 - (-\alpha^2\mu_k));$$

hence its eigenvalues are $n-m$ copies of 0, and $\pm\sqrt{-\alpha^2\mu_k}$ for each eigenvalue μ_k of FE.

Finally

$$p_{A(\alpha)+B(\alpha)}(t) = (t-\alpha)^{n-m} p_{FE}(t^2-\alpha^2) = (t-\alpha)^{n-m} \prod_{k=1}^{m} (t^2 - \alpha^2 - \mu_k);$$

hence its eigenvalues are $n-m$ copies of α, and $\pm\sqrt{\alpha^2+\mu_k}$. □

Lastly, we extend these results to $A(0)$ and $B(0)$.

Theorem 8. *Let* 0 *be an eigenvalue of A of algebraic multiplicity* $n \geq 1$. *Let* $A(0)$, $B(0)$ *be the corresponding big blocks. Then*

$$p_{A(0)B(0)}(t) = p_{A(0)}(t) = t^n,$$
$$p_{A(0)+B(0)}(t) = p_{B(0)}(t).$$

Proof. $A(0)$ is strictly upper triangular, and $B(0)$ is upper triangular. Hence $A(0)B(0)$ is strictly upper triangular, while $A(0)+B(0)$ is upper triangular with the same diagonal entries as $B(0)$. □

Corollary 9. *Let* 0 *be an eigenvalue of A of algebraic multiplicity* $n \geq 1$. *Let* $A(0)$, $B(0)$ *be the corresponding big blocks. Let* $\{\lambda_k\}$ *denote the eigenvalues of* $B(\alpha)$, *with multiplicity. Then the eigenvalues of* $A(\alpha)B(\alpha)$ *are all* 0, *and the eigenvalues of* $A(\alpha)+B(\alpha)$ *are exactly* $\{\lambda_k\}$.

We close by observing that it appears that the diagonalizability hypothesis on $A(\alpha)$ can be weakened or removed entirely, but we are unable to prove this.

Acknowledgment

The authors would like to thank the anonymous referee, who pointed out Theorem 4.4.6 and Lemma 4.4.11 in [Horn and Johnson 1991], which the reader may wish to consult as related to our Theorem 4.

References

[Horn and Johnson 1991] R. A. Horn and C. R. Johnson, *Topics in matrix analysis*, Cambridge Univ. Press, 1991. MR Zbl

[Hrubeš 2016] P. Hrubeš, "On families of anticommuting matrices", *Linear Algebra Appl.* **493** (2016), 494–507. MR Zbl

[Shapiro and Martin 1998] D. B. Shapiro and R. Martin, "Anticommuting matrices: 10456", *Amer. Math. Monthly* **105**:6 (1998), 565–566. MR

[Taussky 1970] O. Taussky, "Sums of squares", *Amer. Math. Monthly* **77** (1970), 805–830. MR Zbl

[Zhong 2017] W. Zhong, "On the eigenvalues of anticommuting matrices", *College Math. J.* **48**:5 (2017), 368–369. MR Zbl

Received: 2020-02-18 Revised: 2020-06-01 Accepted: 2020-06-02

vponomarenko@sdsu.edu *San Diego State University, San Diego, CA, United States*

louisselstad@icloud.com *San Diego State University, San Diego, CA, United States*

INVOLVE 13:4 (2020)
https://doi.org/10.2140/involve.2020.13.633

Combinatorial random knots

Andrew Ducharme and Emily Peters

(Communicated by Kenneth S. Berenhaut)

We explore free knot diagrams, which are projections of knots into the plane which don't record over/under data at crossings. We consider the combinatorial question of which free knot diagrams give which knots and with what probability. Every free knot diagram is proven to produce trefoil knots, and certain simple families of free knots are completely worked out. We make some conjectures (supported by computer-generated data) about bounds on the probability of a knot arising from a fixed free diagram being the unknot, trefoil, or figure-eight knot.

1. Introduction

Knots and links have been objects of mathematical interest for centuries. In 1833, Gauss found the linking integral for two loops. Knots and links were some of the first objects to be studied topologically. They remain a cornerstone of the field of topology, and are useful in many settings beyond their innate two-/three-dimensionalness.

The combinatorial study of knots dates back to Reidemeister, who described a set of three moves which generate all equivalences of knot diagrams. In this article, we investigate a combinatorial aspect of knot theory coming from considering knot diagrams without crossing data, and the knots that result from random assignment of crossing data. Suppose we have a "free knot diagram", as shown in Figure 1. If we randomly assign over/under data to each crossing, how many unknots will we get? For a generic free knot diagram, will any assignment of crossings produce trefoils, figure-eight knots, the knot 5_2, etc.? What is the average crossing number of all knots produced by such assignments to a given free knot diagram?

We began this project by taking an experimental approach: For a set of free knot diagrams with a smallish number of crossings, we computed all the knots which result from assignments of crossing data. We did this using a Mathematica

MSC2010: primary 57M25; secondary 57M27.
Keywords: knot theory, free knot diagrams, random knots, combinatorics.
Supported in part by NSF Grant DMS-1501116.

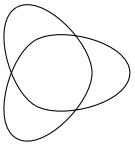


Figure 1. A free knot diagram.

program to calculate the Jones polynomial: though not a perfect invariant, the Jones polynomial can tell apart all prime knots with nine or fewer crossings.

After looking at the data, we made some observations, and in this article prove many of them. As a starting point, we have the following:

Theorem 1.1. *A random knot coming from an n-crossing knot diagram has probability at least $2n/2^n$ of being the unknot.*

This theorem sets a lower bound on the possible amount of unknots, which is intriguing, but does not tell us much information about the vast majority of knots. For example, once we start looking at diagrams with six crossings, this theorem only describes one third of all produced unknots. We want to establish stronger bounds.

Another observation we quickly made was the following:

Theorem 1.2. *Let K be a free knot diagram which is nontrivial. Then K has some assignment of crossings that produces a trefoil diagram.*

We examine four particular categories of free knot diagrams, and describe the knots that result. Computational evidence suggests that some of these knots realize upper or lower bounds on unknots, trefoils, or figure-eight knots.

Conjecture 1.3. *The free 2-braid knot diagram with n crossings realizes the upper bound on trefoils for all sufficiently complicated free diagrams with n and $n+1$ crossings.*

Conjecture 1.4. *The free $(2, n)$-diagram in Figure 2 realizes the lower bound on the trefoils and the upper bound on the unknots for all sufficiently complicated free diagrams with $n+2$ crossings.*

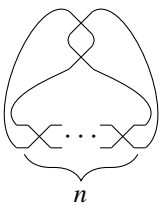


Figure 2. A free $(2, n)$-diagram.

Based on the determined formula for unknots produced by a randomization of a free $(2, n)$-knot, we also propose an absolute maximum of resultant unknot probability coming from a nontrivial free knot diagram, of 0.75.

The structure of this article is as follows. Section 2 establishes notation and defines free knot diagrams. Our general results are in Section 3. The most interesting of these is that every free knot diagram produces trefoils. Sections 4, 5, and 6 compute the complete resolution of 2-braid knot diagrams, the $(2, n)$-tangle knots, and the $(2, 1, n)$-tangle knots, with Section 5 including discussion of the more general (k, n)-tangle knots.

At the end of each section, we include some conjectures which are supported by the data we generated, but do not seem accessible to prove at the moment. Future directions to investigate include these conjectures and working out resolutions of other knot families. It would be great to develop a more theoretical understanding of the role that various structures play in knot resolutions: understanding tangle structure and its role in generating (apparently) minimal-unknot and maximal-unknot examples, and understanding braid structure and proving that the figure-eight knot is universal among prime knot diagrams with braid index 3 and higher.

We thank the referee for calling our attention to [Henrich et al. 2013], which resolves the 2-braid knot diagrams (our Theorem 4.4) and $(2, n)$-tangle knots (Theorem 5.4), and [Cantarella et al. 2017], which, using an immediate consequence of [Taniyama 1989] confirms the existence of the trefoil (Theorem 3.5), figure-eight (Observation 4.9), and the knot 5_2 as resultants of all sufficiently complicated free knot diagrams. We also note [Medina and Salazar 2019; Medina et al. 2019], which contribute a much more systematic understanding of particular knots which must result from free knots. Our discussion of conjectured bounds on unknot, trefoil, and figure-eight probabilities based on the free closures of rational tangles and free composite knots, in addition to our direct proof of Theorem 3.5, appear unique to this work.

2. Background

A knot is a smooth embedding of the circle $\mathbb{S}^1$ into Euclidean three-space $\mathbb{R}^3$, considered up to isotopy. The mathematical study of knots, however, quickly turns into a study of essentially two-dimensional objects as shown in Figure 3. This is a

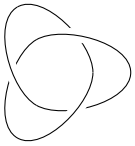


Figure 3. A knot diagram.

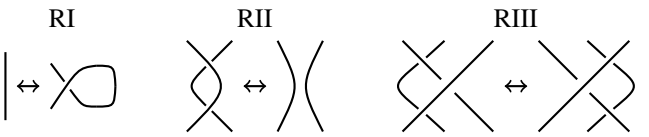


Figure 4. The Reidemeister moves.

knot diagram: a smooth projection of a knot into $\mathbb{R}^2$ in which all crossings have exactly two transverse (not tangent) strands, and the (barely) three-dimensional data of which strand crosses above is denoted by a break in the understand. A *strand* of a knot is the image under the embedding of any interval of the original circle. A *link* is the multicomponent generalization of a knot.

Given two knot diagrams, how do we know if they represent the same knot? We may attempt to directly manipulate one diagram into another. Reidemeister moves, shown in Figure 4, are three isotopies between diagrams which generate all isotopies.

A simple-to-define knot invariant is the *unknotting number*. It is the least number of crossings one can change in any knot diagram of the given knot to unknot it. This change is shown in Figure 5, where the "overstrand" becomes the "understrand" and vice versa. While straightforward to compute for diagrams, the difficulty arises when attempting to show that no other projection of the knot can be unknotted with fewer changes.

The question we consider in this article, of which knots come from which free knot diagrams, is in some sense a converse to the question of unknotting number. Instead of starting with a knot diagram and trying to change it to reach the unknot, we start with only the shape of a knot diagram, and ask where we may land by assigning its crossings.

Random knots have been studied before, for example from a geometric point of view relating to their appearance in random walks and polymers [Dobay et al. 2003]. Some information is known about their average writhe [Diao et al. 2010].

In order to investigate the combinatorial properties of "crossingless" knots, we define a free knot diagram, meaning a projection into the plane of a knot which does not record which strands go over which others at crossings. It is more convenient, however, to say the following:

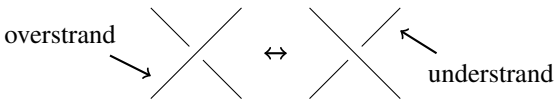
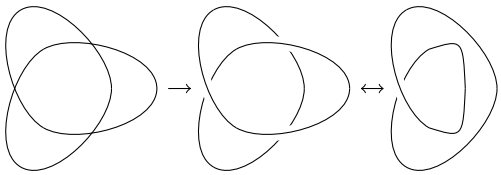


Figure 5. Changing a crossing.

Definition 2.1. A *free knot diagram* is a planar 4-valent graph, considered up to planar isotopy (i.e., continuous deformations in the plane which preserve vertices and edges). The 4-valent vertices are called *free crossings*.

Free knot diagrams have been studied before in [Manturov and Manturov 2010; Manturov 2012; Hanaki 2010] and in [Henrich et al. 2011a; 2011b], where free knot diagrams are called knot shadows (and mixed diagrams, defined below, are called pseudodiagrams).

Example 2.2. Below is a free knot diagram, and an assignment of its crossings which produces the unknot:



Similar simplification with different assignments shows the free trefoil diagram's eight resultants are six unknots and two trefoils.

We may produce a free knot diagram from any knot diagram, by forgetting the over/under information of the crossings:

Definition 2.3. The *shape* of a knot diagram is the free knot diagram which results from making all the crossings of the original knot diagram into free crossings.

Reidemeister moves do not apply to free knot diagrams, since planar isotopies of free knot diagrams preserve crossings. A Reidemeister move (incorrectly) applied to a free knot diagram would create a distinct free knot diagram. We can move back to more familiar ground by "assigning" crossings:

Definition 2.4. When we *assign* a free crossing, we choose an overstrand and an understrand at that crossing. An *assignment* of a free knot is a choice of how to assign each crossing.

Definition 2.5. A *mixed knot diagram* is the projection into the plane of a knot, which records over/under information for only some crossings.

In a mixed knot diagram, we may freely apply Reidemeister moves whenever they involve genuine crossings and no free crossings.

To answer the question of which knots a free knot diagram produces, we assign every possible combination of crossings on the free knot diagram. The two choices at each crossing ensure that for a free knot diagram with n crossings, there are 2^n total resultants. The computation of all resultants creates a sample space of outcomes for a random assignment of crossings for a free knot diagram.

Definition 2.6. The *resultant knot probability* for a given knot R from a free knot diagram F is the probability that the randomization process applied to F will produce R.

Whenever a resultant knot is produced, its mirror image is also produced. Thus, we do not distinguish knots and their mirror images, and the minimum resultant knot probability for an n-crossing free knot diagram is $2/2^n$.

The direct calculation of resultant knot probabilities is intensive, and computation time increases quickly as n increases. In the online supplement, we list the probabilities of the unknot, trefoil, and figure-eight knots, in addition to the expectation value, for the knot shapes of the knots with at most nine crossings. The question of what bounds we can place on resultants in general is explored in the next section.

3. General results

It is well known that any knot diagram can have some of its crossings changed to represent the unknot. The algorithm below for doing so is also well-known; see, e.g., [Adams 2004]. We include this proof as a warm up, and because it uses techniques we will draw on later.

Theorem 3.1. *An n-crossing free knot diagram has a minimum of $2n$ assignments which are the unknot.*

Proof. By an "arc" of a free knot diagram, we mean a portion of the string containing no crossings (that is, a subset which is homeomorphic to an interval).

Choose an arc and an endpoint, and make that point the highest point of the knot (which we now think of as sitting above the page). Now travel away from the arc, and force our path to travel downhill, by making the current strand an overstrand at every unassigned crossing we encounter. When we arrive at the other endpoint of our initial arc, we are at is the lowest point of the knot. We call the arc containing our initial point the "climb"; the rest of the knot is the "downramp".

Currently we are looking down on the knot from above; rotate the knot so that we view it from the side. From here, we see that the knot we created this way is isotopic to the unknot: since every height, other than the very top and bottom points, has exactly two points at that height (one from the climb, the other from the downramp), the identification with a circle is straightforward.

An n-crossing free knot diagram has $2n$ distinct arcs, and as each arc has two endpoints, we have $4n$ distinct climb/downramp pairs on the free knot diagram. But this does not mean we have $4n$ combinatorially distinct unknot diagrams! An easy way to tell apart the unknot diagrams is by their "top track": this is the longest strand containing the climb and only overcrossings. In other words, it stops just before passing through any crossing for the second time. An unknot diagram may

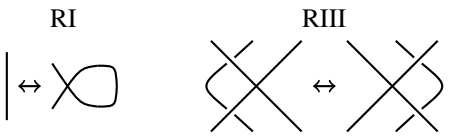


Figure 6. The Reidemeister moves of Theorem 3.2.

share its top track with at most one other unknot diagram — this diagram would traverse the same top track in the opposite direction.

Thus, at least $4n/2 = 2n$ of the 2^n different knot diagrams associated to an n-crossing free knot diagram are diagrams of the unknot. □

Theorem 3.2. *Resultant knot probability for a free or mixed knot diagram is invariant under the free first Reidemeister move and a specific mixed third Reidemeister move* (*shown in Figure 6*).

Proof. These relations hold for either assignment of the free crossing; thus their use does not change resultant knot probability. □

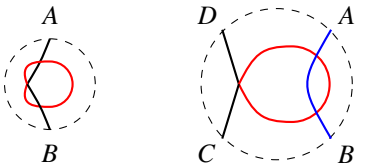
Definition 3.3. A *loop* in a knot diagram or a free knot diagram is a segment of the diagram which starts and ends at the same crossing, and does not cross itself otherwise. The *length* of a loop is the number of crossings that segment passes through (with the crossing of origin/terminus only counting once.) All lengths of loops in knot diagrams are odd numbers.

In later theorems, these results can create essentially trivial counterexamples. For example, when discussing upper bounds, the trefoil, whose unknot probability is 0.75, could have $n-3$ loops of length 1 introduced, thus creating a diagram with n crossings and unknot probability 0.75. To prevent these problems, we define the following:

Definition 3.4. A *minimal* free knot diagram of n crossings is a free knot diagram which contains no loops of length 1 (i.e., those removable via Reidemeister I).

Theorem 3.5. *Let K be a minimal free knot diagram with three or more crossings. Then K has some assignment of crossings making the trefoil knot.*

Proof. Suppose K has a loop of length 3. Zoom in on this loop (drawn in red), which has one of two possible forms:

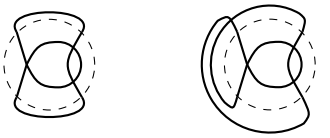


In both cases we assign crossings so that outside of the highlighted disks, the strands of the knot are unknotted. In the first case, along the outside strand, make A the high point, decreasing monotonically to the low point B. At any self-crossing, assign the higher strand to go over the lower strand.

Once A and B are externally trivially connected, it is straightforward to assign crossings to get the trefoil knot: traveling within the disk along the strand from A, assign the crossings so the current strand goes over, then under, then over. The result is a trefoil knot.

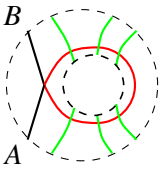
In the second case, there are two subcases to consider. Outside of the highlighted disk, A can either connect to D or C. (If A connected to B, we would have a link, not a knot.) If A connects to D, then we unknot the external strands from each other and themselves by making the A-D and B-C strands downramps and assigning blended crossings so that A-D always goes above B-C. Similarly, if A connects to C, make the A-C and B-D strands downramps, and assign the blended crossings so that A-C always goes above B-D.

Now, it is straightforward to separate and untangle the external strands, so that the knot looks like one of these:



Both are easily made into the trefoil: For the first one, starting at 12 o'clock on the external strand and traveling clockwise, make that strand pass over, then under, then over the other strands it encounters. For the second one, again starting at 12 o'clock on the external strand and traveling clockwise, make that strand pass over, then under, then under, then over the other strands it encounters.

Now suppose K only has loops of length greater than 3. By induction on the length of a loop, we will show that any knot has a resolution that is the trefoil. Zoom in on a single loop:



The inductive step is to assign crossings and perform isotopies to shorten the length of this loop by two. Do this by picking any green strand at the boundary, declaring that location the high point of a downramp which goes through the highlighted disk, and traveling along that strand, assigning every free crossing encountered to have the chosen green strand go over the other strand, until the strand leaves the disk.

The strand thus assigned passes above all other strands on the interior of the disk. Thus, by a combination of Reidemeister II and mixed Reidemeister III moves, it can be isotoped away from the loop, shortening the length of the loop by 2.

In this way, any free knot diagram with a loop of length more than 3 has some mixed knot in its family with a loop of length 3, thus producing a trefoil knot. □

In the following results, we consider connected sums of free knot diagrams. As connected sum is an operation on knots, and its well-definedness on knot diagrams relies on the Reidemeister moves, we shouldn't expect it to be well-defined on free knots. However, if we are concerned with which knots arise as assignments of a given free knot, then there is no problem: once crossings are assigned, Reidemeister moves are allowed. Thus, in Theorem 3.6 through Consequence 3.17, we simply refer to a connected sum without worrying about which one is meant.

Theorem 3.6. *Suppose K_1 and K_2 are knots with shapes S_1 and S_2 that are components of a connected sum $S_1\#S_2$. Then the probability of getting an unknot resultant U from $S_1\#S_2$ is $P(S_1 \to U)P(S_2 \to U)$.*

Proof. The unknot is a prime knot and can only be created by the knot sum of two unknots; thus S_1 and S_2 must be simultaneously, and independently, assigned to produce the unknot. □

This simple product of probabilities ensures composites must respect any upper bound on unknot probability that its components are subject to.

Theorem 3.7. *Suppose K_1 and K_2 are knots with shapes S_1 and S_2 that are components of a connected sum $S_1\#S_2$. Then the probability of getting a resultant nontrivial prime K_3 from $S_1\#S_2$ is*

$$P(S_1 \to U)P(S_2 \to K_3) + P(S_1 \to K_3)P(S_2 \to U).$$

Proof. If we assign the crossings of S_2 such that it becomes the unknot, then we are left with the free knot diagram of S_1, from which we can produce K_3. Thus, the probability of K_3 is at least

$$P(S_1 \to U)P(S_2 \to K_3) + P(S_1 \to K_3)P(S_2 \to U).$$

Further K_3-knots could only arise if some connected sum of K_1 and K_2 or any of their resultants could create K_3. The assumption that K_3 is prime makes this impossible, so the earlier sum calculates the resultant knot probability for K_3. □

Theorem 3.8. *Suppose K_1 and K_2 are prime knots with shapes S_1 and S_2 that are components of a connected sum $S_1\#S_2$. Then the probability of getting a resultant composite knot $K_3\#K_4$ from $S_1\#S_2$ is*

$$\begin{aligned} P(S_1 \to U)P(S_2 \to K_3\#K_4) + P(S_1 \to K_3\#K_4)P(S_2 \to U) \\ + P(S_1 \to K_3)P(S_2 \to K_4) + P(S_1 \to K_4)P(S_2 \to K_3). \end{aligned}$$

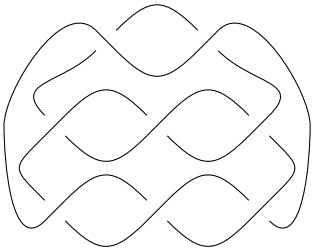


Figure 7. The knot 8_5.

Proof. This is a straightforward probability computation. It is worth mentioning that it *is* possible to get composite resultants from shapes of prime knots; for example, the connect sum of two trefoils arises from the shape of knot 8_5, shown in Figure 7. □

Definition 3.9. A *recursive sum* S^N of a knot K with shape S is a connected sum of N copies of S.

Theorem 3.10. *Let K_1 be a prime knot with a shape S. Then the resultant knot probability for a nontrivial prime K_2, $P(S^{N+1} \to K_2)$, is*

$$P(S \to K_2)P(S \to U)^N + P(S \to U)P(S^N \to K_2).$$

Proof. Given $S^{N+1} = S^N \# S$, we apply the formula from Theorem 3.7 and get

$$P(S^N \to U)P(S \to K_2) + P(S^N \to K_2)P(S \to U).$$

Applying Theorem 3.6 N times shows $P(S^N \to U) = P(S \to U)^N$. □

Definition 3.11. Using the notation $x_K = P(S^K \to K_2)$, $\alpha = P(S \to U)$, and $\beta = P(S \to K_2)$, we rewrite the recurrence relation for $P(S^{N+1} \to K_2)$ as

$$x_{N+1} = \alpha^N \beta + \alpha x_N.$$

The appropriate initial condition, for a generic shape S and resultant K, is $P(S^0 \to K) = 0$, as a crossingless loop, the identity of the connect sum operation, has no nontrivial resultants.

Theorem 3.12. *Let K_1 be a prime knot with shape S such that a prime K_2 is a resultant ($\beta \neq 0$, and thus $\alpha < 1$). Then the sequence of resultant knot probability of K_2 from the recursive sum of S, $P(S^N \to K_2)$, strictly increases, then has a maximum value at two or fewer steps after which the probability strictly decreases for every additional connected sum.*

Proof. Suppose that for some N, $x_N \geq x_{N+1}$. Then let us compare x_{N+2} and x_{N+1}:

$$x_{N+2} = \alpha^{N+1}\beta + \alpha x_{N+1} < \alpha^N \beta + \alpha x_N = x_{N+1},$$

with the inequality following from $\alpha < 1$ and $x_{N+1} \leq x_N$. Therefore, it is possible that the sequence increases for initial values, but once equality or a decrease is achieved, the sequence strictly decreases for all further indices. □

Corollary 3.13. *Let K_1 be a prime knot with shape S. Then the resultant knot probability of K_2 from the recursive sum of S, $P(S^N \to K_2)$, has a maximum value at $N = 1$ if and only if $P(S \to U) \leq 0.5$.*

Proof. As $x_1 = \beta$ and $x_2 = \alpha^1\beta + \alpha x_1 = 2\alpha\beta$, we know $x_1 \geq x_2$ if and only if $\frac{1}{2} \geq \alpha$. According to Theorem 3.12, x_1 is then a maximum. □

Assuming our upper bounds hold over the prime knots, Corollary 3.13 forces recursive sums whose base knot have a resultant unknot percentage less than or equal to 0.5 to similarly respect our proposed absolute bounds over the primes for the trefoil and figure-eight knots.

Theorem 3.14. *Let K_1 be a nontrivial prime knot with shape S. Then the resultant knot probability of K_2 from the recursive sum of S, $P(S^N \to K_2)$, has a limit $\lim_{N\to\infty} P(S^N \to K_2) = 0$.*

Proof. According to Theorem 3.12, after some number of steps, x_N is monotone decreasing. As a probability, x_N is also bounded below by zero; therefore the sequence has a limit L. Note the nontriviality of K_1 forces $0 < \alpha < 1$. Then

$$L = \lim_{N\to\infty} \alpha^N\beta + \lim_{N\to\infty} \alpha x_N = 0 + \alpha L.$$

This can only be true if $L = 0$. □

Theorem 3.15. *Let K_1 be a prime knot with shape S. Then for a nontrivial prime K_2 the resultant knot probability is*

$$P(S^N \to K_2) = N \cdot P(S \to U)^{N-1} P(S \to K_2).$$

Proof. We proceed by induction on N. Note $0\alpha^{-1}\beta = 0 = x_0$, and $1\alpha^0\beta = \beta = x_1$, showing agreement with our recursive formula for $P(S_1^N \to K_2)$ in the base cases.

Now suppose $x_k = k\alpha^{k-1}\beta$ for some $N = 0, 1, \ldots, k$. Then $x_{k+1} = \alpha^{k+1-1}\beta + \alpha x_k$. By the inductive hypothesis,

$$x_{k+1} = \alpha^k\beta + \alpha(k\alpha^{k-1}\beta) = \alpha^k\beta + k\alpha^k\beta = (k+1)\alpha^k\beta$$

and thus the formula holds for all N. □

Theorem 3.16. *Let K_1 be a nontrivial prime knot with shape S with resultant unknot probability α. Then the resultant knot probability for a nontrivial prime K_2*

has a singular maximum at

$$\begin{aligned} N &= 1 \quad \text{if } \alpha < \tfrac{1}{2}, \\ N &= 2 \quad \text{if } \tfrac{1}{2} < \alpha < \tfrac{2}{3}, \\ N &= 3 \quad \text{if } \tfrac{2}{3} < \alpha < \tfrac{3}{4}. \end{aligned}$$

Proof. If the resultant knot probability for K_2 has solely one maximum value after N recursive sums, then the difference between the N-th and surrounding steps gives the inequalities

$$\begin{aligned} (N+1)\alpha^N\beta - N\alpha^{N-1}\beta < 0, \\ N\alpha^{N-1}\beta - (N-1)\alpha^{N-2}\beta > 0. \end{aligned}$$

Solving each for α places bounds on what the resultant unknot probability can be while having a maximum solely at particular values of N:

$$\frac{N-1}{N} < \alpha < \frac{N}{N+1}.$$

Choosing low values of N then produce the ranges above. Note if α is equivalent to one of these bounds, both N and $N+1$ are (equivalent) maxima. □

Assuming Conjecture 1.4 holds, its corollary that the absolute maximum resultant unknot probability is 0.75 forces the following consequence of the conjecture.

Consequence 3.17. *All resultant knot probabilities of nontrivial prime knots from recursive sums have a maximum in the first four connected sums.*

Proof. As forced by Theorem 3.12, the resultant knot probability will attain a maximum at two adjacent steps if the difference between each is equal to zero ($N \geq 1$), or, specifically, if

$$(N+1)\alpha^N\beta - N\alpha^{N-1}\beta = 0 \quad \text{or} \quad (N+1)\alpha - N = 0.$$

Solving for N, the first step where the maximum occurs, as a function of α, we get

$$N(\alpha) = \frac{\alpha}{1-\alpha}.$$

The derivative of $N(\alpha)$, $1/(1-\alpha)^2$, is positive for all α, so the largest final step containing a maximum will occur for the largest possible α, which we conjecture to be $\alpha = 0.75$. Since $N(0.75) = 3$, the final step containing a maximum will occur at $N = 4$, as the value of the resultant probability stays the same. □

Recursive sums (and other connect sums) have the potential to create counterexamples to our later proposed bounds on resultant trefoil and figure-eight probabilities. In those cases, we note that they apply to solely prime knots. These recursive sum counterexamples depend more on the unknot probability of the base knot shape

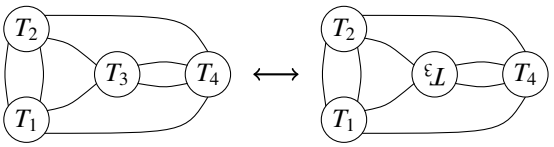


Figure 8. Two free knot diagrams with the same resultant knot probabilities.

than the base shape's resultant probability. For example, 7_1 appears to produce the most trefoils at a rate of 32.8125%, but its recursive sum only produces a higher percentage (approximately 35.89%) for $(7_1)^2$. Meanwhile the recursive sum of the trefoil, with its absolute maximum resultant unknot probability of 75%, produces conjecture breaking results for $(3_1)^2$ through $(3_1)^6$.

We present two final general conjectures on resultant probabilities, based on observations of all free knot diagrams with up to nine crossings:

Conjecture 3.18. *For all free knot diagrams and nontrivial knots, the resultant knot probability is less than* 0.5.

There appear to be further manipulations we can perform within free knot diagrams which do not change their resultant knot probability. To describe these, we need the language of tangles.

Definition 3.19. A *tangle* is a portion of a knot or link contained in a circle which intersects the boundary at exactly four points. Reidemeister moves and planar isotopies are allowed in the interior of the circle.

A free tangle is a tangle with free crossings instead of true crossings. Observations show that 90° rotations and reflections across the diagonals of a tangle (i.e., an axis connecting two opposite anchor points) change the resultant knot probability. However, knot mutations [Viro 1976; Conway 1970], which are vertical and horizontal reflections as well as 180° rotations, seem not to change resultant knot probability.

Conjecture 3.20. *A knot and its mutants have the same resultant knot probability.*

For example, the free knot diagrams shown in Figure 8 with a variety of choices of T_i (including nonsymmetric possibilities) were seen to have the same resultant knot probabilities.

4. Tangles and the n_1-knots

Recall that the notation for tangles introduced in [Conway 1970] gives us concise descriptors for a large family of knots. It is useful not only because of its brevity, but because it calculates an invariant of tangles, their continued fraction.

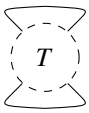


Figure 9. A tangle.

Theorem 4.1 [Conway 1970]. *Two rational tangles are isotopic to each other if their continued fraction is equivalent.*

A tangle, shown in Figure 9, is made into a knot or link by taking its closure, which connects the upper and lower two pairs of exterior arcs. The continued fraction continues to differentiate between knots once we move to closures of tangles by the following process:

Lemma 4.2 (from [Schubert 1956], as quoted in [Kauffman and Lambropoulou 2002]). *Suppose there exist two rational tangles with continued fractions p/q and p'/q', where p, q and p', q' are relatively prime. Let $K(p/q)$ and $K(p'/q')$ be the knots formed by the closures of the respective rational tangles. Then $K(p/q)$ and $K(p'/q')$ are equivalent* (*up to isotopy*) *if and only if*

- $p = p'$,
- *either* $q \equiv q' \bmod p$ *or* $qq' \equiv 1 \bmod p$.

Definition 4.3. The n_1-knots are the closures of two-strand braids with n crossings (n is odd). These knots include the trefoil, pentafoil, 7_1, 9_1, and the unknot (with the first Reidemeister relation applied).

An odd number of crossings is required since an even number of crossings would require two components, making the result a link. The name n_1 comes from their position as the first entry in knot tabulations of knots with n crossings.

Theorem 4.4. *The free n_1-knot produces $\binom{n}{(n-k)/2}$ left k_1-knots and $\binom{n}{(n-k)/2}$ right k_1-knots ($k \le n$).*

Proof. Previously, we have counted a knot and its reflection in the same category, even if it is chiral. Here, we count a knot and its reflection as distinct knots.

A braid can have two types of crossings, "positive slope" and "negative slope", referring to the slope of the overstrand. Given a free n_1-knot, we may choose ℓ of its crossings to be positive slope crossings and the remaining $n - \ell$ to be negative slope crossings. Then, if $0 < \ell < n$, we may find a positive crossing next to a negative crossing, and remove the pair via Reidemeister II. Proceeding thus until all crossings of one type have been removed, we are left with either $n - 2\ell$ or $2\ell - n$ crossings of a single type. Substituting $\ell = (n - k)/2$ or $\ell = (n + k)/2$ produces the formula stated above. □

Corollary 4.5. *The expected number of crossings for a resultant of an n_1-knot is $2^{1-n}\sum_{k=3}^{n} k\binom{n}{(n-k)/2}$, where the values of k are odd.*

Proof. Since free n_1-knots produce an equal number of right and left k_1-knots, the probability of getting a resultant with k crossings is $\frac{2}{2^n}\binom{n}{(n-k)/2}$. The free 2-braid knot diagram with one crossing is the unknot; thus its probability has a coefficient of zero in the expectation value. Then the expectation value sum, over the allowed odd crossing values, starts at the trefoil and goes up to the n_1-knot. □

Theorem 4.6. *For fixed odd k, the limit as the number of crossings n goes to infinity of the resultant k_1-knot probability is 0 for free n_1-knots.*

Proof. Applying Stirling's approximation to the resultant k_1-knot probability leads to

$$\frac{2}{2^n}\binom{n}{(n-k)/2} \sim \sqrt{\frac{8}{\pi}}\frac{n-k}{(n+k)^2}\frac{n^{n+1/2}}{(n-k)^{n/2}(n+k)^{n/2}}.$$

The highest power of n occurring in the denominator is $n+2$, and is larger than the highest power of n in the numerator $\left(\text{which is } n+\frac{3}{2}\right)$. Thus the resultant k_1-knot probability goes to zero as n goes to infinity. □

Conjecture 4.7. *Among all minimal prime free knot diagrams with n and $n+1$ crossings, where n is odd and $n \geq 3$, the free n_1-knot has the most trefoil descendants.*

To create a trefoil, an assignment of crossings must (after Reidemeister II cancellations) result in three alternating crossings in a row. When trying to find assignments that create such a shape, the n_1-knots should then have an advantage.

Conjecture 4.7 would also allows us to strengthen Theorem 4.6's results by applying it to all knots, forcing the limit of the resultant k_1-knot probability as the number of crossings n goes to infinity to be 0.

Consequence 4.8. *For prime knots, the absolute maximum k_1-knot probability is $2^{1-k^2}\binom{k^2}{(k^2-k)/2}$, and consequently, the maximum trefoil percentage is* 32.8125.

Proof. We create a sequence g_n calculating the resultant k_1-knot probability for n_1-knots, where n must be incremented by 2:

$$g_n = \frac{2\binom{n}{(n-k)/2}}{2^n} = \frac{1}{2^{n-1}}\binom{n}{(n-k)/2}.$$

The next term in the sequence can be written as

$$g_{n+2} = \frac{(n+2)(n+1)}{(n+2-k)(n+2+k)}g_n.$$

This sequence will be monotonically decreasing when $n > k^2$ and monotonically increasing when $n < k^2 - 2$. Observing that $g_{k^2-2} = g_{k^2}$, we have two occurrences

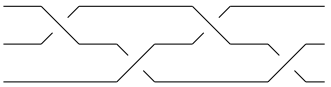


Figure 10. The minimal braid representation of the figure-eight knot.

of the maximum probability for a given k_1-knot:

$$2^{3-k^2}\binom{k^2-2}{(k^2-k-2)/2} = 2^{1-k^2}\binom{k^2}{(k^2-k)/2}.$$

Then for the trefoil ($k = 3$), 7_1 and 9_1 have the maximum trefoil percentage of 32.8125. □

Observation 4.9. *Let K be a minimal knot with four or more crossings, which is neither an n_1-knot nor a connect sum of n_1-knots. Then K has some assignment of crossings making the figure-eight knot.*

The above statement is called an observation, not a conjecture, because it is proven in [Cantarella et al. 2017] using results from [Taniyama 1989]. However, it still seems desirable to us to have a more direct proof of this, perhaps along the lines of the proof of Theorem 3.5 above.

Theorem 4.4 confirms that all resultants of free n_1-knots are also 2-braid knot diagrams, so no figure-eights are produced. These appear to be the only free knots without a figure-eight resultant. This is sensible since they are the only knots with braid index of 2.

The figure-eight knot has the minimal braid representation shown in Figure 10. If one of the outer strands of a minimal braid diagram was connected to its neighboring internal strand by only one crossing, the external strand could be removed via braid stabilization, so a braid index greater than 3 forces the existence of more than four free crossings. Then seemingly there should be a route to unknotting some remaining crossings to produce a figure-eight resultant.

Based on our experimental evidence, we propose slightly stronger upper bounds on the unknot probability as this would require four nontrivial resultants of a given free knot diagram.

Conjecture 4.10. *The resultant figure-eight probability is less than or equal to the resultant trefoil probability for a free knot diagram coming from algebraic tangles.*

The hypothesis concerning algebraic tangles is included because we know of exactly one counterexample to this conjecture: the knot 9_{40}, which is not the closure of an algebraic tangle. It has 66 resultant trefoils and 78 resultant figure-eights. Yet this conjecture holds for all other free knots with nine or fewer crossings.

Consequence 4.11. *For all algebraic free knot diagrams with n and $n+1$ crossings, where n is odd and $n \geq 3$, the trefoil probability of the n_1-knot is the upper bound on figure-eight probability.*

Proof. This extends the previous results from trefoils to figure-eights using the above conjecture over all algebraic diagrams. □

The largest resultant figure-eight percentage for all knots through eight crossings is shared by 7_7 and 8_{12} at 15.625%, so there is likely a more restrictive bound to be found.

Conjecture 4.12. *The absolute maximum resultant figure-eight percentage is* 15.625%.

When moving beyond the algebraic knots, this number appears to remain the upper bound, as 9_{40} and 9_{47} have respective resultant figure-eight percentages of 15.234% and 13.281%.

5. The (k, n)-knots

Definition 5.1. The $(2, n)$-knots (Figure 11, left), closures of the $(2, n)$-tangles, include the trefoil, the figure-eight, 5_2, 6_1, 7_2, 8_1, 9_2, and so on.

One may think about generating these knots by repeatedly twisting a loop, and then coupling the two ends together. This construction method demonstrates that these knots should produce a lot of unknots! Any time the upper two strands can be separated, every assignment of the remaining crossings must give an unknot.

Definition 5.2. The (k, n)-knots (Figure 11, right) are the generalization of the $(2, n)$-knots, and include knots like 7_3 and 8_3. One of k or n must be even or else we get a link, not a knot, from the closure.

Using continued fractions, Lemma 4.2 shows $(2, n)$-knots are nontrivial if $n \geq 1$, and distinct for distinct n. It also tells us the (k, n)-knots are distinct for $k \geq 3,\ n \geq 3$.

Theorem 5.3. *The resultant knot probability of a free (k, n)-knot is the same as the resultant knot probability for a free (n, k)-knot.*

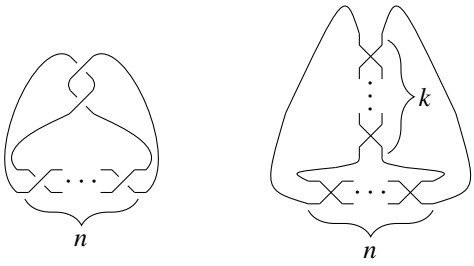


Figure 11. A $(2, n)$-knot (left) and a (k, n)-knot (right).

unknot		$(2, k)$-knot	$(2, (k-1))$-knot
n even	n odd	n, k same parity	
$2^{n+1}+2\binom{n}{n/2}$	$2^{n+1}+2\binom{n}{(n-1)/2}$	$2\binom{n}{(n-k)/2}$	

Table 1. The complete resultants of the free $(2, n)$-knot, $k \leq n$.

Proof. Begin with a free (k, n)-knot as depicted in Figure 11. Via a spherical isotopy, bring the far right strand around the rest of the free knot diagram so it is now to the left of everything else. Next isotope the resulting diagram so the k-tangle is made horizontal by a 90° clockwise rotation, and the n-tangle is made vertical by a 270° clockwise rotation. The result is the (n, k)-free knot diagram. □

In this and the following section, we present many theorems without proofs, when those proofs are straightforward computations, most of which rely chiefly on the results of the explication of the n_1-knot within each free knot structure.

Theorem 5.4. *The free $(2, n)$-knot's complete resultants $(k \leq n)$ are given in Table 1. (To calculate the resultant probability for an even/odd resultant with ℓ crossings when both n and k are odd/even, let $k = \ell + 1$.)*

Theorem 5.5. *The number of unknots produced by a (k, n)-free knot diagram with k, n both even is*

$$2^k\binom{n}{n/2}+2^n\binom{k}{k/2}-\binom{n}{n/2}\binom{k}{k/2}$$

and with k odd and n even is

$$2^k\binom{n}{n/2}+2\binom{k}{(k-1)/2}\binom{n}{(n-2)/2}.$$

Corollary 5.6. *The expected number of crossings for a resultant of a $(2, n)$-knot is*

$$2^{-(n+1)}\sum_{k=2}^{n}(2k+3)\binom{n}{(n-k)/2},$$

where the values of k and n are even, and

$$2^{-(n+1)}\left(3\binom{n}{(n-1)/2}+\sum_{k=3}^{n}(2k+3)\binom{n}{(n-k)/2}\right),$$

where the values of k and n are odd.

Theorem 5.7. *The limit of resultant unknot probability as the number of crossings n $(k \leq n)$ goes to infinity is 0.5 for $(2, n)$-knots.*

Proof. The resultant unknot probabilities for a $(2, n)$-knot are

$$\frac{1}{2}+\frac{1}{2^{n+1}}\binom{n}{n/2}, \quad \frac{1}{2}+\frac{1}{2^{n+1}}\binom{n}{(n-1)/2}$$

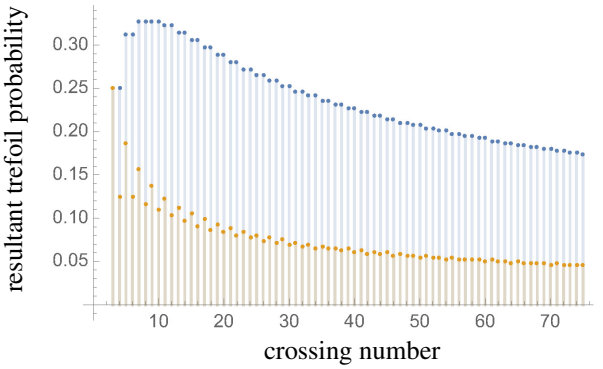


Figure 12. Resultant trefoil probabilities for the n_1-knots (blue) and $(2, n)$-knots (orange).

when n is even and odd, respectively. Applying the asymptotic behavior in Theorem 4.6 for $k = 0$ and $k = 1$ to the second terms, only the $\frac{1}{2}$ term remains for both odd and even n. □

Conjecture 5.8. *The free $(2, n)$-knots realize the upper bound on the unknots for minimal free knot diagrams with $n + 2$ crossings.*

Consequence 5.9. *The absolute maximum value for the unknot percentage from a nontrivial knot is 75%.*

Proof. Similar to Consequence 4.8, we define infinite series e_n and o_n equivalent to the probability of the unknot for the $(2, n)$-knots for even and odd n:

$$e_n = \frac{1}{2} + \frac{1}{2^{n+1}}\binom{n}{n/2}, \quad o_n = \frac{1}{2} + \frac{1}{2^{n+1}}\binom{n}{(n-1)/2}.$$

Both series are easily seen to be monotonically decreasing, so the (2, 1)-knot — i.e., the trefoil — has the maximum probability, at 75%. □

Conjecture 5.10. *The free $(2, n)$-knots are the lower bound for trefoils for minimal prime algebraic free knot diagrams with $n + 2$ crossings.*

As so many assignments of the free $(2, n)$-knot are forced to be the unknot, this does not leave much room for trefoils. If true, the free $(2, n)$-knots are not the only knots that realize the lower bound. For example, 7_2 shares the same resultant trefoil percentage, 15.625%, with 7_4 and 7_7. Again, we insert the algebraic knot qualifier for the lone counterexample of 9_{40}.

Between the two trefoil bounding conjectures, the space of possible trefoil probabilities would be known. Figure 12 shows these possible ranges of probabilities for free knots with up to 75 crossings in between those for the n_1-knots (blue) and $(2, n)$-knots (orange).

6. The (2, 1, n)-knots

Definition 6.1. The $(2, 1, n)$-knots are again the closure of the $(2, 1, n)$-tangle when n is odd; see Figure 13. Examples include the figure-eight, 6_2, and 8_2.

In applying this section's formulas, if a term's lower integer index for its binomial coefficient (the k in the "n choose k" language) is negative, that term should be understood to be ignored.

Theorem 6.2. *A free* $(2, 1, n)$*-knot produces* $12\binom{n}{(n-1)/2} + 2\binom{n}{(n-3)/2}$ *unknots.*

Proof. We break into eight cases, depending on the assignment of the crossings in the $(2, 1)$-portion of the tangle, and then the computation is very similar to the computations above. □

Theorem 6.3. *A free* $(2, 1, n)$*-knot produces*

$$8\binom{n}{(n-3)/2} + 2\binom{n}{(n-1)/2} + 2\binom{n}{(n-5)/2}$$

trefoils, *and*, *in general*,

$$8\binom{n}{(n-k)/2} + 2\binom{n}{(n-k+2)/2} + 2\binom{n}{(n-k-2)/2}$$

k_1*-knots* ($k \le n+2$).

Proof. First consider the different possible assignments of the $(2, 1)$-tangles shown in Figure 14.

The first four assignments (red) of the free $(2, 1, n)$-tangle produce only closures of two-strand braids, so we get $4 \cdot 2\binom{n}{(n-3)/2}$ trefoils. In the green assignments, a ± 2-tangle can be added to a ∓ 5-tangle or a ± 1-tangle to produce a trefoil: there are $2\binom{n}{(n-5)/2}$ ways to make the free n-tangle into a ∓ 5-tangle, and $2\binom{n}{(n-1)/2}$ ways to make the free n-tangle into a ± 1-tangle.

If one's goal is to get a k-tangle, there are $4 \cdot 2\binom{n}{(n-k)/2}$ ways to get this from the red assignments, and $2\binom{n}{(n-k-2)/2} + 2\binom{n}{(n-k+2)/2}$ ways from the green.

The final two assignments (blue) are the closures of the $(\pm 2, \pm 1, m)$-tangles, which have continued fractions

$$m + \frac{1}{\pm 1 \pm \frac{1}{2}} = m \pm \tfrac{2}{3} = \frac{3m \pm 2}{3}.$$

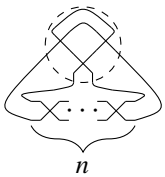


Figure 13. A $(2, 1, n)$-knot.

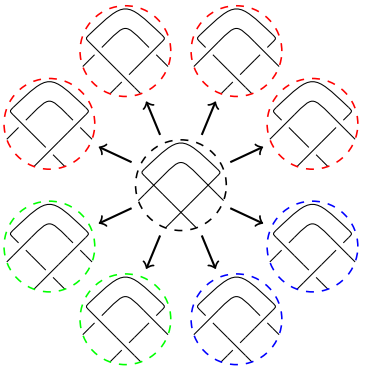


Figure 14. Possible arrangements of (2, 1)-tangles.

We appeal to Lemma 4.2, as any k_1-knot would be $K(k/1)$, where k is an odd integer, so $q = 1$ and $q' = 3$. Thus these assignments cannot produce any k_1-knots, as the difference of 3 and 1 is even, so the odd k ensures $1 \not\equiv 3 \bmod k$ and $3 \not\equiv 1 \pmod k$. □

Theorem 6.4. *A free* $(2, 1, n)$*-knot produces* $2\binom{n}{(n-1)/2}$ *figure-eights, and* $2\binom{n}{(n-k)/2}$ *of* $(2, 1, k)$*-knots,* $k \le n$.

Proof. As in the proof above, we focus on the resolution of the (2, 1)-tangle. The red and green assignments are all k_1-knots, so any figure-eight knot must come from a blue assignment. This is then a straightforward combinatorics exercise. □

Theorem 6.5. *A free* $(2, 1, n)$*-knot produces* $2\binom{n}{(n-k)/2}$ $(3, k-1)$*-knots for* $3 \le k \le n$, k *odd.*

Proof. The only remaining unexamined combination of this knot's component tangles are when $(\pm 2, \pm 1)$-tangles are connected to $\mp k$-tangles $(3 \le k \le n)$. Without loss of generality, we'll examine the $(-2, -1, k)$-case with continued fraction $\frac{3k-2}{3}$. Now using Lemma 4.2 in a positive sense, we can identify the closures of this tangle as isotopic to the closures of $(3, \ell)$-tangles $\left(\text{whose continued fraction is } \ell + \frac{1}{3} = \frac{3\ell+1}{3}\right)$ since their numerators are equal when $\ell = k - 1$, and their identical denominators trivially satisfy the second condition by reflexivity. Since the (2, 1)-tangles are already assigned, the number of each resultant formed depends on the number of ways to assign the n-tangle to be k crossings long, namely $2\binom{n}{(n-k)/2}$ ways. □

This completes the elucidation of the $(2, 1, n)$-knot's resultants.

References

[Adams 2004] C. C. Adams, *The knot book: an elementary introduction to the mathematical theory of knots*, Amer. Math. Soc., Providence, RI, 2004. MR Zbl

[Cantarella et al. 2017] J. Cantarella, A. Henrich, E. Magness, O. O'Keefe, K. Perez, E. Rawdon, and B. Zimmer, "Knot fertility and lineage", *J. Knot Theory Ramifications* **26**:13 (2017), art. id. 1750093. MR Zbl

[Conway 1970] J. H. Conway, "An enumeration of knots and links, and some of their algebraic properties", pp. 329–358 in *Computational Problems in Abstract Algebra* (Oxford, 1967), Pergamon, Oxford, 1970. MR Zbl

[Diao et al. 2010] Y. Diao, C. Ernst, K. Hinson, and U. Ziegler, "The mean squared writhe of alternating random knot diagrams", *J. Phys. A* **43**:49 (2010), art. id. 495202. MR Zbl

[Dobay et al. 2003] A. Dobay, J. Dubochet, K. Millett, P.-E. Sottas, and A. Stasiak, "Scaling behavior of random knots", *Proc. Natl. Acad. Sci. USA* **100**:10 (2003), 5611–5615. MR Zbl

[Hanaki 2010] R. Hanaki, "Pseudo diagrams of knots, links and spatial graphs", *Osaka J. Math.* **47**:3 (2010), 863–883. MR Zbl

[Henrich et al. 2011a] A. Henrich, N. MacNaughton, S. Narayan, O. Pechenik, R. Silversmith, and J. Townsend, "A midsummer knot's dream", *College Math. J.* **42**:2 (2011), 126–134. MR Zbl

[Henrich et al. 2011b] A. Henrich, N. Macnaughton, S. Narayan, O. Pechenik, and J. Townsend, "Classical and virtual pseudodiagram theory and new bounds on unknotting numbers and genus", *J. Knot Theory Ramifications* **20**:4 (2011), 625–650. MR Zbl

[Henrich et al. 2013] A. Henrich, R. Hoberg, S. Jablan, L. Johnson, E. Minten, and L. Radović, "The theory of pseudoknots", *J. Knot Theory Ramifications* **22**:7 (2013), art. id. 1350032. MR Zbl

[Kauffman and Lambropoulou 2002] L. H. Kauffman and S. Lambropoulou, "Classifying and applying rational knots and rational tangles", pp. 223–259 in *Physical knots: knotting, linking, and folding geometric objects in* $\mathbb{R}^3$ (Las Vegas, NV, 2001), edited by J. A. Calvo et al., Contemp. Math. **304**, Amer. Math. Soc., Providence, RI, 2002. MR Zbl

[Manturov 2012] V. O. Manturov, "Free knots and parity", pp. 321–345 in *Introductory lectures on knot theory*, edited by L. H. Kauffman et al., Ser. Knots Everything **46**, World Sci. Publ., Hackensack, NJ, 2012. MR Zbl

[Manturov and Manturov 2010] O. V. Manturov and V. O. Manturov, "Free knots and groups", *J. Knot Theory Ramifications* **19**:2 (2010), 181–186. MR Zbl

[Medina and Salazar 2019] C. Medina and G. Salazar, "The knots that lie above all shadows", *Topology Appl.* **268** (2019), art. id. 106922. MR Zbl

[Medina et al. 2019] C. Medina, J. Ramírez-Alfonsín, and G. Salazar, "On the number of unknot diagrams", *SIAM J. Discrete Math.* **33**:1 (2019), 306–326. MR Zbl

[Schubert 1956] H. Schubert, "Knoten mit zwei Brücken", *Math. Z.* **65** (1956), 133–170. MR Zbl

[Taniyama 1989] K. Taniyama, "A partial order of knots", *Tokyo J. Math.* **12**:1 (1989), 205–229. MR Zbl

[Viro 1976] O. J. Viro, "Nonprojecting isotopies and knots with homeomorphic coverings", *Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov.* (*LOMI*) **66** (1976), 133–147. In Russian; translated in *J. Soviet Math.* **12**:1 (1979), 86–96. MR Zbl

Received: 2020-02-24 Revised: 2020-06-30 Accepted: 2020-07-07

ducharmean@gmail.com *Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL, United States*

epeters3@luc.edu *Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL, United States*

INVOLVE 13:4 (2020)
https://doi.org/10.2140/involve.2020.13.655

Conjugation diameter of the symmetric groups

Assaf Libman and Charlotte Tarry

(Communicated by Kenneth S. Berenhaut)

The conjugation diameter of a group G is the largest diameter of its Cayley graphs with respect to conjugation-invariant generating sets. It is a strong form of the extensively studied concept of the diameter of G. We compute the conjugation diameter of the symmetric groups.

1. Introduction and main results

Let G be a finite group. Let $\operatorname{diam}(G, S)$ denote the diameter of the associated Cayley graph $\Gamma(G, S)$ with respect to a generating set S. Set $\operatorname{diam}(G) = \sup\{\operatorname{diam}\Gamma(G, S)\}$, where the supremum is taken over all generating sets S. This concept has been studied for several decades and was the subject of intensive activity; see [Babai et al. 1990], which gives a good survey. Particular attention was given to the diameter of the symmetric groups [Babai and Seress 1992; Helfgott and Seress 2014] due to its relevance in computing science and networks [Preparata and Vuillemin 1981].

In this note we study the *conjugation diameter* of a group G, which we denote by $\Delta(G)$. That is, $\Delta(G) = \sup\{\operatorname{diam}\Gamma(G, S)\}$, where S runs through all generating sets which are *conjugation-invariant and conjugation-finite*, i.e., unions of finitely many conjugacy classes in G. Conjugation diameter has been studied under the name C-width by Bardakov, Tolstykh and Vershinin [Bardakov et al. 2012].

Kędra, Martin and the first author had a geometric motivation in studying conjugation diameter. Any generating set S gives rise to a word norm on G, namely the minimum length of a word in $S \cup S^{-1}$ needed to express an element of G. Then $\operatorname{diam}(G, S)$ is the diameter of G with respect to this norm and is a measure of the "efficiency" S generates. If S is conjugation-invariant then so is the associated word norm. Conjugation-invariant norms were studied by Burago, Ivanov and Polterovich [Burago et al. 2008], who introduced the concept of bounded groups, namely groups for which every conjugation-invariant norm has finite diameter. In [Kędra et al. 2018] Kędra, Martin and the first author gave several refinements of this concept for groups G which are finitely normally generated; namely there exists a finite

MSC2010: 05E15, 20B30.
Keywords: conjugation diameter, symmetric groups.

$X \subseteq G$ such that $\langle\langle X \rangle\rangle = G$. These refinements are defined by the diameter of G with respect to conjugation-invariant word norm and are therefore related to $\Delta(G)$.

For example, it is shown in [Kędra et al. 2018, Theorem 6.3] that all noncompact connected semisimple Lie groups G are uniformly bounded, namely $\Delta(G) < \infty$. In fact (unpublished notes) it can be shown that $\Delta(\mathrm{SL}(2, \mathbb{R})) = 4$ and $\Delta(\mathrm{PSL}(2, \mathbb{R})) = 3$ and $\Delta(\mathrm{SL}(2, \mathbb{C})) = 3$ and $\Delta(\mathrm{PSL}(2, \mathbb{C})) = 2$. The second author showed in [Tarry 2020, Chapter 7] that $\Delta(\mathrm{PSL}(n, \mathbb{C})) \leq 6(n-1)$ for all $n \geq 3$. If R is a principal ideal domain with only $d < \infty$ maximal ideals then $\Delta(\mathrm{PSL}(n, R)) \leq 12d(n-1)$ for any $n \geq 3$ [Kędra et al. 2018, Theorem 6.3].

In general, calculating $\Delta(G)$ is difficult and the purpose of this note is to compute this invariant for some finite groups. If G is finite abelian then $\Delta(G) = \mathrm{diam}(G)$, which was calculated in [Klopsch and Lev 2003], where they showed that if $G = C_{n_1} \times \cdots \times C_{n_r}$ is the canonical decomposition [Rotman 1973, Corollary 4.7], where $n_1 \mid \cdots \mid n_r$, then $\Delta(G) = \sum_i \lfloor n_i/2 \rfloor$. Here $\lfloor x \rfloor$ is the floor of x.

Beyond abelian groups calculations are more involved. Let $p < q$ be distinct primes such that $p \mid (q-1)$ and let G be the unique nonabelian group of order pq. An easy application of Sylow's theorems gives the following theorem, which should be compared with [Babai and Seress 1992, Proposition 5.5], where it is shown that $\mathrm{diam}(G) < 3q$.

Theorem 1.1. *Let $p < q$ be primes and G a nonabelian group of order pq. Then*

$$\Delta(G) = \max\left\{\tfrac{p-1}{2}, 2\right\}.$$

The main result of this paper is the calculation of the conjugation diameter of the symmetric groups. It should be compared with the celebrated results in [Helfgott and Seress 2014].

Theorem 1.2. *Let S_n denote the symmetric group, $n \geq 2$. Then*

$$\Delta(S_n) = n - 1.$$

2. Norms and conjugation diameter

Let X be a subset of a group G. Set $X^{-1} = \{x^{-1} : x \in X\}$. If $X, Y \subseteq G$ set $XY = \{xy : x \in X, y \in Y\}$ and let X^n denote $X \cdots X \subseteq G$ (n factors).

Definition 2.1. Let X be a subset of a group G. Set $\mathrm{ccs}(X) = \{gxg^{-1} : x \in X, g \in G\}$, the union of the conjugacy classes of the elements of X. For any $n \geq 0$ define subsets $B_X(n)$ of G as follows. Set

$$B_X(0) = \{1\} \quad \text{and} \quad B_X(1) = \{1\} \cup \mathrm{ccs}(X) \cup \mathrm{ccs}(X^{-1}).$$

For any $n \geq 1$ set

$$B_X(n) = B_X(1)^n \subseteq G.$$

If $X = \{g\}$ is a singleton, we will often write $B_g(n)$.

Thus, $B_X(n)$ is the set of all "words" of length at most n in the conjugates of the elements of X and their inverses. The following proposition follows directly from the definitions. See [Kędra et al. 2018, Lemma 2.3] and [Tarry 2020, Lemma 1.15] for details.

Proposition 2.2. *Let X, Y be subsets of G:*

(i) *$B_X(n)$ is closed under conjugation in G.*

(ii) *If $X \subseteq Y$ then $B_X(n) \subseteq B_Y(n)$ for all $n \geq 0$.*

(iii) *$B_X(m) \cdot B_X(n) = B_X(m+n)$.*

(iv) *If $Y \subseteq B_X(n)$ for some $n \geq 0$ then $B_Y(m) \subseteq B_X(mn)$ for all $m \geq 0$.*

Definition 2.3. We say that $X \subseteq G$ *normally generates* G if $G = \langle\langle X \rangle\rangle$. We say that G is *finitely normally generated* if it contains a finite normally generating set.

Note that X normally generates G if and only if $\bigcup_{n\geq 0} B_X(n) = G$. Thus, the following definition makes sense (the minimum is taken over a nonempty set of integers).

Definition 2.4. Suppose that X normally generates G. Define $\|\cdot\|_X : G \to \mathbb{R}$ by

$$\|g\|_X = \min\{n \geq 0 : g \in B_X(n)\}.$$

Clearly $\|\cdot\|_X$ is a conjugation-invariant norm on G [Tarry 2020, Proposition 1.19]. We define

$$\|G\|_X = \operatorname{diam}(G, \|\cdot\|_X) = \sup\{\|g\|_X : g \in G\}.$$

It is immediate from the definitions that

$$\|G\|_X = \inf\{n : G \subseteq B_X(n)\}. \tag{1}$$

In particular if $X \subseteq Y$ normally generate G then $\|G\|_Y \leq \|G\|_X$. Clearly, $B_X(n)$ is the closed ball of radius n centred at $1 \in G$ with respect to the metric $\|\cdot\|_X$ induces on G.

Definition 2.5. The *conjugation diameter* of a finitely normally generated group G is

$$\Delta(G) = \sup\{\|G\|_X : X \subseteq G \text{ normally generates } G \text{ and } |X| < \infty\}.$$

We call G *uniformly bounded* if $\Delta(G) < \infty$; see [Kędra et al. 2018, Definition 2.6].

3. pq-groups

Proof of Theorem 1.1. Let Q be a Sylow q-subgroups of G. Then $Q \trianglelefteq G$ since $p < q$. Since G is not abelian, no Sylow p-subgroup of G can be normal and no element of G has order pq.

Our first goal is to prove that any $g \in G$ of order p normally generates G and

$$\|G\|_g = \max\{2, \tfrac{p-1}{2}\}. \tag{2}$$

Let $C_G(g)$ be the centraliser of g. Then either $|C_G(g)| = p$ or $|C_G(g)| = pq$ since $g \in C_G(g)$. The latter is impossible since it implies that $\langle g \rangle$ is a central Sylow p-subgroup of G. We deduce that $|C_G(g)| = p$ and therefore

$$|\mathrm{ccs}(g)| = [G : C_G(g)] = q.$$

Consider the quotient homomorphism $\pi : G \to G/Q$. Since $G/Q \cong C_p$ is abelian, π must be constant on conjugacy classes of G. Then $\mathrm{ccs}(g) \subseteq gQ$ since $\pi(\mathrm{ccs}(g)) = \bar{g}$ and equality holds since they have the same cardinality.

By Definition 2.1, and since $Q \trianglelefteq G$,

$$B_g(1) = \{1\} \cup gQ \cup g^{-1}Q.$$

Since $Q \nsubseteq B_g(1)$, it follows that $\|G\|_g > 1$. Also, $gQ \cdot g^{-1}Q = Q$, which implies that $B_g(2) = g^{-2}Q \cup g^{-1}Q \cup Q \cup gQ \cup g^2Q$. Using induction one shows that

$$B_g(n) = \bigcup_{k=-n}^{n} g^k Q, \quad n \geq 2.$$

Now, $\langle g \rangle$ is a Sylow p-subgroup of G and its elements form a complete set of representatives for the cosets of Q. If $p = 2$ or $p = 3$ then $\langle g \rangle = \{1, g^{\pm 1}\}$ and therefore $B_g(2) = G$ so $\|G\|_g = 2$. If $p > 3$ then p is odd and $\frac{p-1}{2} \geq 2$ and

$$\langle g \rangle = \{g^k : -\tfrac{p-1}{2} \leq k \leq \tfrac{p-1}{2}\}.$$

Therefore $B_g\big(\frac{p-1}{2}\big) = G$ and $B_g(n) \neq G$ if $n < \frac{p-1}{2}$. It follows that $\|G\|_g = \frac{p-1}{2}$ in this case and we have established (2). In particular

$$\Delta(G) \geq \max\{\tfrac{p-1}{2}, 2\}.$$

Let $X \subseteq G$ be any normally generating subset of G. No element of order pq exists and if all elements of X have order q then $\langle\langle X \rangle\rangle = Q \trianglelefteq G$, a contradiction. So there exists $g \in X$ of order p and we have seen that g normally generates G and

$$\|G\|_X \leq \|G\|_g = \max\{\tfrac{p-1}{2}, 2\}.$$

It follows that $\Delta(G) \leq \max\{\frac{p-1}{2}, 2\}$ and equality holds. □

4. The symmetric groups

4.1. *Notation and basic facts.* Conjugation of elements $g, h \in G$ is denoted by

$$g^h = hgh^{-1}.$$

Any $\sigma \in S_n$ can be written as a product of disjoint cycles of lengths $k_1, \ldots, k_r$, where $k_i \geq 1$ and $\sum_i k_i \leq n$. We call σ a $(k_1, \ldots, k_n)$-*cycle*. Cycle structure determines the conjugacy class [Hall 1959, Theorem 5.13] and we denote the conjugacy class of σ by

$$[k_1, \ldots, k_m].$$

Conjugation of a k-cycle $(i_1 \cdots i_k)$ by $\tau \in S_n$ is the k-cycle $(\tau(i_1) \cdots \tau(i_k))$ [Rotman 1973, Lemma 3.9]. The inverse of a k-cycle is a k-cycle and hence any $\sigma \in S_n$ is conjugate to σ^{-1}.

Let $\mathrm{fix}(\sigma)$ denote the set of fixed points and $\mathrm{supp}(\sigma)$ denote the support. If $\mathrm{fix}(\sigma)$ is not empty then σ is conjugate to $\sigma' \in S_{n-1}$.

Lemma 4.2. *Consider $\tau \in S_n$. Then $B_\tau(n)$ is the set of elements of the form $\tau^{\lambda_1} \cdots \tau^{\lambda_\ell}$, with conjugation by $\lambda_1, \ldots, \lambda_\ell \in S_n$, where $\ell \leq n$.*

Proof. The elements of $B_\tau(n)$ are products of at most n conjugates of $\tau^{\pm 1}$. Since τ^{-1} is conjugate to τ the result follows. □

Lemma 4.3. *Suppose that $\tau \in S_n$ is a product $\tau = \alpha\beta$ of permutations with disjoint supports, where $\alpha \in S_k$ and $\beta \in S_{n-k}$ for some k. Then $B_\tau(2)$ contains all elements of the form $\alpha^{\lambda_1}\alpha^{\lambda_2}$ for any $\lambda_1, \lambda_2 \in S_k$.*

Proof. Choose $\theta \in S_{n-k}$ such that $\beta^\theta = \beta^{-1}$. Then $\tau^{\lambda_1}\tau^{\lambda_2\theta} = \alpha^{\lambda_1}\alpha^{\lambda_2}\beta\beta^\theta = \alpha^{\lambda_1}\alpha^{\lambda_2}$. □

Lemma 4.4. *Let $n \geq 2$:*

(i) *If $X \subseteq S_n$ normally generates S_n then X contains an odd permutation.*

(ii) *Conversely, any odd permutation normally generates S_n.*

Proof. (i) If X contains only even permutations then $\langle\langle X \rangle\rangle \subseteq A_n \trianglelefteq Sn$.

(ii) The only proper normal subgroups of S_n are A_n and the Klein group $K \subseteq A_4$ if $n = 4$. □

Obtaining a lower bound for $\Delta(S_n)$ is easy.

Proposition 4.5. *Let $\tau \in S_n$ be a transposition. Then τ normally generates S_n and $\|S_n\|_\tau = n - 1$.*

Proof. Any permutation is a product of 2-cycles, so τ normally generates. Any k-cycle is a product of $k-1$ transpositions; see [Rotman 1973, Proof of Theorem 3.4]. Hence any $\sigma \in [k_1, \ldots, k_m]$ is a product of $\sum_i k_i - m \leq n - m \leq n - 1$ transpositions. A product of m transpositions has at least $n - m$ orbits showing that an n-cycle cannot be written as a product of less than $n - 1$ transpositions. This shows that $\|S_n\|_\tau = n - 1$. □

Corollary 4.6. *We have $\Delta(S_n) \geq n - 1$ for any $n \geq 2$.*

Our goal now is to compute $\Delta(S_n)$. A major role will be played by 3-cycles and $(2,2)$-cycles. An important feature they have is that we can obtain them "cheaply" from any nonidentity permutation.

Lemma 4.7. *Let $\tau \in S_n$ be a nonidentity element where $n \geq 4$. Then*:

(i) *$B_\tau(2)$ contains a 3-cycle if either τ is a transposition or if τ contains a cycle of length ≥ 3.*

(ii) *$B_\tau(2)$ contains a $(2,2)$-cycle if τ is a transposition or it contains a $(2,2)$-cycle or it contains a cycle of length ≥ 4.*

Proof. If τ is a transposition then $(1\,2)(2\,3) = (1\,2\,3)$ and $(1\,2)(3\,4)$ give the result. In the other cases, the calculations

$$(1\,2)(3\,4)\cdot(1\,3)(2\,4) = (1\,4)(2\,3),$$
$$(1\,2\,3\cdots k)\cdot(k\,k-1\cdots 3\,1\,2) = (1\,3\,2), \qquad k \geq 3,$$
$$(1\,2\,3\,4\cdots k)\cdot(k\,k-1\cdots 4\,1\,2\,3) = (1\,3)(2\,4), \quad k \geq 4,$$

together with Lemma 4.3, give the result. □

4.8. The next two propositions tell us that, with some fine print, by multiplying a 3-cycle τ with a permutation σ we may either

(a) split one of the cycles of σ into three disjoint parts, or

(b) fuse two disjoint cycles in σ and split the result in two, or

(c) fuse three cycles of σ into one cycle.

Clearly operations (a) and (c) are inverse of each other and the operation (b) is inverse to itself. Similarly, subject to some fine print, by multiplying a $(2,2)$-cycle τ with σ we may either

(a) split one of the cycles of σ into three disjoint cycles, or

(b1) split two cycles of σ into two cycles each or,

(b2) fuse two cycles of σ and split the result into two cycles, or

(c1) fuse three cycles of σ, or

(c2) fuse two cycles and split a third, or

(d) fuse two pairs of disjoint cycles.

Thus, 3-cycles and $(2,2)$-cycles provide us with a variety of "operations" on conjugacy classes in S_n.

The following calculations are left for the reader:

$$(1\,2\cdots m)\cdot(i\;j) = (1\cdots i\;\, j+1\cdots m)(i+1\cdots j), \quad 1 \leq i < j \leq m, \tag{3}$$
$$(1\,2\cdots \ell)(\ell+1\cdots m)\cdot(\ell\,m) = (1\,2\cdots m), \qquad 1 \leq \ell < m. \tag{4}$$

Proposition 4.9. *Let $C = [k_1, \dots, k_r]$ be a conjugacy class in S_n, where $k_i \geq 1$ and $\sum_i k_i \leq n$. Then $C \cdot [3]$ contains the following conjugacy classes in S_n, where $p', p'', p''' \geq 1$:*

(a) $[p', p'', p''', k_2, \dots, k_r]$, *where* $p' + p'' + p''' = k_1 \geq 3$.

(b) $[p', p'', k_3, \dots, k_r]$, *where* $r \geq 2,\ k_1 \geq 2,\ p' + p'' = k_1 + k_2$ *and* $p' \neq k_1$.

(c) $[k_1 + k_2 + k_3, k_4, \dots, k_r]$, *where* $r \geq 3$.

Proof. (a) Set $p = k_1$. Consider $1 < j < i \leq p$ (notice that $p \geq 3$). By inspection

$$(1\,2 \cdots p) \cdot (1\,i\,j) = (1\,i+1 \cdots p)(2 \cdots j)(j+1 \cdots i).$$

If $p' + p' + p''' = k_1$, set $j = p' + 1$ and $i = p' + p'' + 1$, and check that the resulting permutation belongs to $[p''', p'', p']$.

(b) Set $p = k_1$ and $q = k_2$. For any $i \neq p, p+q$ we have

$$(i\ p\ p+q) = (p\ p+q)(i\ p+q),$$

so (4) and (3) imply

$$(1\,2 \cdots p)(p+1 \cdots p+q) \cdot (i\ p\ p+q) = (1\,2 \cdots i)(i+1 \cdots p+q)$$

is a product of cycles of length i and $p + q - i$.

(c) Set $p = k_1$ and $q = k_2$ and $t = k_3$. By inspection

$$(1 \cdots p)(p+1 \cdots p+q)(p+q+1 \cdots p+q+t) \cdot (p\ p+q\ p+q+t) = (1\,2 \cdots p+q+t). \quad \square$$

Proposition 4.10. *Let $C = [k_1, \dots, k_r]$ be a conjugacy class in S_n, where $k_i \geq 1$ and $\sum_i k_i \leq n$. Then $C \cdot [2, 2]$ contains the following conjugacy classes in S_n, where $p', p'', p''', q', q'' \geq 1$:*

(a) $[p', p'', p''', k_2, \dots, k_r]$, *where* $p''' \geq 2$ *and* $p' + p'' + p''' = k_1 \geq 4$.

(b1) $[p', p'', q', q'', k_3, \dots, k_r]$, *where* $r \geq 2$ *and* $p' + p'' = k_1 \geq 2$.

(b2) $[p', p'', k_3, \dots, k_r]$, *where* $r \geq 2,\ p' + p'' = k_1 + k_2 \geq 4,\ p' \leq k_1 - 2$, *and if* $k_1 \geq 3$ *and* $k_2 \geq 2$ *then* $p' \leq k_1 - 1$.

(c1) $[k_1 + k_2 + k_3, k_4, \dots, k_r]$, *where* $r \geq 3$ *and* $k_1 \geq 2$.

(c2) $[p' + p'', k_2 + k_3, k_4, \dots, k_r]$, *where* $r \geq 3$ *and* $p' + p'' = k_1 \geq 2$.

(d) $[k_1 + k_2, k_3 + k_4, k_5, \dots, k_r]$, *where* $r \geq 4$.

Proof. (a) Set $p = k_1$. Choose some $1 < i < j < p$ (notice that $p \geq 4$). By (3)

$$(1\,2 \cdots p) \cdot (1\,i)(j\ p) = (1\,i+1 \cdots j)(2 \cdots i)(j+1 \cdots p).$$

By choosing $i = p' + 1$ and $j = p' + p'''$, we obtain a (p''', p', p'')-cycle.

(b1) Set $p = k_1$ and $q = k_2$. Choose $1 \le i < j \le p$ such that $j - i = p'$ and $p+1 \le k < m \le p+q$ such that $m - k = q'$ and apply (3) to

$$(1\,2\cdots p)(p+1\cdots p{+}q)\cdot(i\ j)(k\,m).$$

(b2) Set $p = k_1$ and $q = k_2$. Choose $1 \le i < j \le p+q$ distinct from p, $p+q$ (notice that $p+q \ge 4$ by assumption). By (4) and (3)

$$(1\,2\cdots p)(p+1\cdots p{+}q)\cdot(p\ p{+}q)(i\ j) = (1\,2\cdots p{+}q)\cdot(i\ j)$$

is a product of two cycles of lengths $j-i$ and $p+q-j+i$. If we choose $i = 1$ and $2 \le j \le p-1$ we obtain a $(p', p{+}q{-}p')$-cycle for any $1 \le p' \le p-2$. If $p \ge 3$ and $q \ge 2$ we may choose $i = 2$ and $j = p+1$ to get a $(p{-}1, q{+}1)$-cycle.

(c1) Set $p = k_1 \ge 2$ and $q = k_2$ and $t = k_3$. Check that

$$(1\cdots p)(p+1\cdots p{+}q)(p{+}q{+}1\cdots p{+}q{+}t)\cdot(p\ p{+}q)(1\ p{+}q{+}t)$$

is a $p{+}q{+}t$-cycle (use (4)).

(c2) Set $p = k_1$ and $q = k_2$ and $t = k_3$. For any $1 \le i < p$ (note that $p \ge 2$)

$$(1\cdots p)(p+1\cdots p{+}q)(p{+}q{+}1\cdots p{+}q{+}t)\cdot(i\ p)(p{+}q\ p{+}q{+}t)$$

is a $(i, p{-}i, q{+}t)$-cycle (use (3) and (4)).

(d) If $\alpha_1\alpha_2\beta_1\beta_2$ is a product of disjoint cycles (possibly of length 1), use (4) twice to get an $\alpha\beta$ product of disjoint cycles of lengths $|\alpha_1| + |\alpha_2|$ and $|\beta_1| + |\beta_2|$. □

Notation 4.11. In light of the discussion in 4.8, the cases of Proposition 4.9 will be referred to O_3(a), O_3(b) and O_3(c) and those of Proposition 4.10 as O_2(a), O_2(b1), O_2(b2) etc. This reminds us that we view 3-cycles and (2, 2)-cycles as "operations" on permutations which either split or fuse cycles.

Lemma 4.12. *Consider $\sigma \in S_n$ with cycle structure $[k_1, \ldots, k_r]$, where $k_i \ge 1$ and $\sum_i k_i = n$. Let $1 \le m \le n$. Then there exist $\ell \ge 0$ and 3-cycles $\alpha_1, \ldots, \alpha_\ell$ such that $r \ge 2\ell + 1$ and if we set $\tilde{k} = \sum_{i=1}^{2\ell+1} k_i$ then the cycle structure of $\sigma\alpha_1\cdots\alpha_\ell$ is either*

(i) *$[\tilde{k}, k_{2\ell+2}]$, where $r = 2\ell+2$ and $\tilde{k} \le m-1$, or*

(ii) *$[\tilde{k}, k_{2\ell+2}, \ldots, k_r]$ and $\tilde{k} \ge m$ and $\sum_{i=1}^{2\ell-1} k_i < m$.*

In fact, for any $0 \le j \le \ell$ the cycle structure of $\sigma\alpha_1\cdots\alpha_j$ is

$$\left[\sum_{i=1}^{2j+1} k_i, k_{2j+2}, \ldots, k_r\right].$$

(*Notice that in* (ii) *it may happen that $r = 2\ell+1$; hence $\tilde{k} = n$ and $\sigma\alpha_1\cdots\alpha_{2\ell+1}$ is an n-cycle*).

Proof. Apply O_3(c) repeatedly to choose 3-cycles $\alpha_1, \ldots, \alpha_\ell$ that "fuse" the first cycle with the next two until the first instance when $\sum_{i=1}^{2\ell+1} k_i \geq m$ or until $\sigma\alpha_1 \cdots \alpha_\ell$ contains only one or two cycles (If there are three or more cycles left and $\sum_{i=1}^{2\ell+1} k_i < m$, we will proceed applying O_3(c)). In the first two cases we have established (ii) (since $\sum_i k_i = n \geq m$) and in the third case (two cycles remaining) it is (i). □

Proposition 4.13. *Let $\tau \in S_n$ be an odd permutation. Suppose that τ contains a k-cycle, where $k \geq 3$ is odd and that $n - k \geq 2$. Then*

$$\|S_n\|_\tau \leq \Delta(S_{n-k}) + k.$$

Proof. By assumption $n \geq k + 2 \geq 5$. Let $\sigma \in S_n$ be a nonidentity element. Our goal is to prove that $\|\sigma\|_\tau \leq \Delta(S_{n-k}) + k$. We will do this in three steps.

Step I: There are 3-cycles $\alpha_1, \ldots, \alpha_t$, where $t \leq \frac{k-1}{2}$ such that $\sigma\alpha_1 \cdots \alpha_t$ contains a k-cycle.

Proof of Step I. Let $[k_1, \ldots, k_r]$ be the cycle structure of σ, where $\sum_i k_i = n$ and $k_1 \geq \cdots \geq k_r$. Note that $k_1 \geq 2$ since $\sigma \neq \mathrm{id}$. Recall Notation 4.11.

If σ is a transposition, its cycle structure is $[1, \ldots, 1, 2]$, with $n - 2 \geq k$ fixed points. Apply O_3(c) repeatedly $t = \frac{k-1}{2}$ times with 3-cycles $\alpha_1, \ldots, \alpha_t$ to obtain $\sigma\alpha_1 \cdots \alpha_t \in [k, 1, \ldots, 1, 2]$ and we are done.

Assume that σ is not a transposition. Then either $k_1 \geq 3$ or $k_1, k_2 \geq 2$. Hence, if $r \geq 2$ then $k_1 + k_2 \geq 4$.

Use Lemma 4.12 with $m = k$ to find 3-cycles $\alpha_1, \ldots, \alpha_\ell$ such that $\ell \geq 0$ and $r \geq 2\ell + 1$ and if we set $\tilde{k} = \sum_{i=1}^{2\ell+1} k_i$ then the cycle structure of $\xi = \sigma\alpha_1 \cdots \alpha_\ell$ is either

(i) $[\tilde{k}, k_{2\ell+2}]$, where $\tilde{k} \leq k - 1$, or

(ii) $[\tilde{k}, k_{2\ell+2}, \ldots, k_r]$, where $\tilde{k} \geq k$ and $\sum_{i=1}^{2\ell-1} k_i < k$.

Case (i): We have $\xi \in [\tilde{k}, n - \tilde{k}]$, where $\tilde{k} < k$. Use O_3(b) to find a 3-cycle $\alpha_{\ell+1}$ such that $\xi\alpha_{\ell+1} \in [k, n-k]$ contains a k-cycle. It remains to show that $\ell + 1 \leq \frac{k-1}{2}$. If $\ell = 0$ then we are done. If $\ell \geq 1$ then $r \geq 3$ and

$$k_1 + k_2 + \sum_{i=3}^{2\ell+1} k_i = \sum_{i=1}^{2\ell+1} k_i \leq k - 1.$$

Since $k_1 + k_2 \geq 3$ and $k_i \geq 1$ we get $3 + 2\ell - 1 \leq k - 1$ so $\ell \leq \frac{k-3}{2}$ and we are done.

Case (ii): If $\tilde{k} = k$ then ξ contains a k-cycle. If $\tilde{k} = k+1$ then $r \geq 2\ell+2$ since $n > k+1$ and we use O_3(b) to find a 3-cycle $\alpha_{\ell+1}$ such that $\xi\alpha_{\ell+1} \in [k, \tilde{k} + k_{2\ell+2} - k, \ldots, k_r]$ contains a k-cycle. If $\tilde{k} \geq k + 2$, use O_3(a) to find a 3-cycle $\alpha_{\ell+1}$ such that $\xi\alpha_{\ell+1} \in [k, 1, \tilde{k} - k - 1, \ldots]$ contains a k-cycle. Thus, a product of σ with at most $\ell + 1$ 3-cycles gives a permutation which contains a k-cycle.

If $\ell = 0$ then $\ell + 1 = 1 \leq \frac{k-1}{2}$ and we are done. If $\ell \geq 2$ then $r \geq 5$ and

$$\sum_{i=1}^{2\ell-1} k_i = k_1 + k_2 + \sum_{i=3}^{2\ell-1} k_i \geq 4 + (2\ell - 3).$$

By assumption $\sum_{i=1}^{2\ell-1} k_i \leq k - 1$ so $\ell \leq \frac{k-2}{2}$ and since k is odd, $\ell \leq \frac{k-3}{2}$. Therefore $\ell + 1 \leq \frac{k-1}{2}$ and we are done.

It remains to consider the case $\ell = 1$. If $k \geq 5$ then $\ell + 1 \leq \frac{k-1}{2}$ and we are done. Assume $k = 3$. By assumption $k_1 = \sum_{i=1}^{2\ell-1} k_i \leq k - 1 = 2$. Then $k_2 \leq k_1 \leq 2$ and since σ is not a transposition, $k_2 = 2$; namely $\sigma \in [2, 2, \dots]$. Use O_3(b) to replace α_1 with a 3-cycle so that $\sigma\alpha_1 \in [3, 1, \dots]$ and we are done $\left(\text{since } 1 \leq \frac{k-1}{2}\right)$. This completes the proof of Step I. □

Step II: If $\mu \in S_n$ contains a k-cycle then $\|\mu\|_\tau \leq \Delta(S_{n-k}) + 1$.

Proof of Step II. Write $\mu = \mu_0\mu_k$ as a product of disjoint permutations, where $\mu_k \in S_k$ is a k-cycle and $\mu_0 \in S_{n-k}$. By assumption $n - k \geq 2$. Similarly, $\tau = \tau_0\tau_k$. Since τ is an odd permutation and τ_k is an even permutation (a cycle of odd length), τ_0 is an odd permutation in S_{n-k} and by Lemma 4.4 it normally generates it. By Lemma 4.2 there are $\lambda_1, \dots, \lambda_\ell \in S_{n-k}$, where $\ell \leq \Delta(S_{n-k})$, such that

$$\mu_0 = \tau_0^{\lambda_1} \cdots \tau_0^{\lambda_\ell}.$$

Choose $\theta \in S_k$ such that $\tau_k^\theta = \tau_k^{-1}$. Since k is odd, τ_k^2 is a k-cycle, so there is $\pi \in S_k$ such that $\tau_k^\pi = \tau_k^2$.

If ℓ is odd then

$$\tau^{\lambda_1}\tau^{\lambda_2\theta}\tau^{\lambda_3}\tau^{\lambda_4\theta} \cdots \tau^{\lambda_{\ell-2}}\tau^{\lambda_{\ell-1}\theta}\tau^{\lambda_\ell} = (\tau_0^{\lambda_1} \cdots \tau_0^{\lambda_\ell}) \cdot ((\tau_k\tau_k^\theta)^{(\ell-1)/2} \cdot \tau_k) = \mu_0\tau_k$$

is conjugate to μ (since both μ_k and τ_k are k-cycles) so $\|\mu\|_\tau \leq \ell \leq \Delta(S_{n-k})$.

Assume that ℓ is even. If $\ell = 0$ then $\mu = \mu_k$ is a k-cycle. Choose some $\epsilon \in S_{n-k}$ such that $\tau_0^\epsilon = \tau_0^{-1}$ and then

$$\tau^\epsilon \cdot \tau = (\tau_0^\epsilon\tau_0) \cdot (\tau_k^2) = \tau_k^2$$

is a k-cycle, and hence is conjugate to μ. Now, $n - k \geq 2$ so $\|\mu\|_\tau = 2 \leq \Delta(S_{n-k}) + 1$ by Corollary 4.6 as needed. If $\ell \geq 2$ then

$$\begin{aligned}\tau^{\lambda_1\pi}\tau^{\lambda_2\theta}\tau^{\lambda_3}\tau^{\lambda_4\theta}\tau^{\lambda_5} \cdots \tau^{\lambda_{\ell-1}}\tau^{\lambda_\ell\theta} &= (\tau_0^{\lambda_1} \cdots \tau_0^{\lambda_\ell}) \cdot (\tau_k^\pi\tau_k^\theta\tau_k\tau_k^\theta \cdots \tau_k\tau_k^\theta) \\ &= \mu_0 \cdot (\tau_k^2\tau_k^{-1} \cdot (\tau_k\tau_k^{-1})^{(\ell-2)/2}) = \mu_0\tau_k\end{aligned}$$

is conjugate to μ so $\|\mu\|_\tau \leq \ell \leq \Delta(S_{n-k})$. This completes the proof of Step II. □

Step III: We show that $\|\sigma\|_\tau \leq \Delta(S_{n-k}) + k$.

Proof of Step III. By Step I there are 3-cycles $\alpha_1, \dots, \alpha_t$, where $t \le \frac{k-1}{2}$, such that $\mu = \sigma\alpha_1 \cdots \alpha_t$ contains a k-cycle. By Step II, $\|\mu\|_\tau \le \Delta(S_{n-k}) + 1$. Since τ contains a k-cycle of length $k \ge 3$, Lemma 4.7 shows that $\|\alpha_i\|_\tau \le 2$. Therefore

$$\|\sigma\|_\tau \le \|\mu\|_\tau + \sum_{i=1}^{t} \|\alpha_i^{-1}\|_\tau \le \Delta(S_{n-k}) + 1 + 2t \le \Delta(S_{n-k}) + k. \qquad \square$$

There are three $(2,2)$-cycles in S_4, and the product of any two is equal to the third. Therefore if τ is a $(2,2)$-cycle in S_n there exists $\pi \in S_n$ such that $\mathrm{supp}(\pi) \subseteq \mathrm{supp}(\tau)$ and $\tau^\pi \tau$ is a $(2,2)$-cycle.

Proposition 4.14. *Let $\tau \in S_n$ be an odd permutation, $n \ge 7$. Suppose that τ contains a (p,q)-cycle, where $p \ge q \ge 2$ are even and $n - (p+q) \ge p$. Then*

$$\|S_n\|_\tau \le \Delta(S_{n-(p+q)}) + p + q.$$

Proof. We will prove that if $1 \ne \sigma \in S_n$ then $\|\sigma\|_\tau \le \Delta(S_{n-(p+q)}) + p + q$. Throughout the proof the cycle structure of σ is $[k_1, \dots, k_r]$ such that $k_i \ge 1$ and $\sum_i k_i = n$ and $k_1 \ge \cdots \ge k_r$. Recall Notation 4.11.

Step I: There exist $\alpha_1, \dots, \alpha_t \in S_n$ such that $t \le \frac{p+q-2}{2}$ and such that $\xi = \sigma\alpha_1 \cdots \alpha_t$ contains a (p,q)-cycle and the following hold. If $p = 2$ then every α_i is a $(2,2)$-cycle and if $p \ge 4$ then each α_i is either a 3-cycle or a $(2,2)$-cycle.

Proof of Step I. Assume first that $p = 2$. Hence, $q = 2$. If $k_1 \ge 5$ then use O_2(a) to find a $(2,2)$-cycle α_1 such that $\sigma\alpha_1 \in [2, k_1 - 4, 2, \dots]$, i.e., $\sigma\alpha_1$ contains a $(2,2)$-cycle, and we are done (since $t = 1 \le \frac{p+q-2}{2} = 1$).

If $k_1, k_2 \ge 3$ then use O_2(b1) to find a $(2,2)$-cycle α_1 such that $\sigma\alpha_1 \in [2, k_1 - 2, 2, k_2 - 2, \dots]$ and we are done.

If $k_1 = 4$ and $k_2 = 2$ then use O_2(b1) to find a $(2,2)$-cycle α_1 such that $\sigma\alpha_1 \in [2, 2, 1, 1, \dots]$. If $k_1 = 4$ and $k_2 = 1$ then $r \ge 3$ and $k_3 = 1$ (since $n \ge 7$) and we use O_2(c2) to get $\sigma\alpha_1 \in [2, 2, 2, \dots]$ and we are done.

Suppose that $k_1 = 3$ and $k_2 = 2$. Then $r \ge 3$ since $n \ge 7$. If $k_3 = 2$ then σ contains a $(2,2)$-cycle and we are done. Otherwise $k_3 = 1$. Then $r \ge 4$ since $n \ge 7$ and $\sigma \in [3, 2, 1, 1, \dots] = [3, 1, 1, 2, \dots]$ and we use O_2(c2) to find a $(2,2)$-cycle α_1 such that $\sigma\alpha_1 \in [2, 1, 2, 2, \dots]$ and we are done.

If $k_1 = 3$ and $k_2 = 1$ then $r \ge 4$ and $k_3 = 1$ and use O_2(c2) to get $\sigma\alpha_1 \in [2, 1, 2, \dots]$ and we are done.

If $k_1 = 2$ and $k_2 = 2$ then σ contains a $(2,2)$-cycle we are done. If $k_1 = 2$ and $k_2 = 1$ then σ is a transposition which fixes at least four points (since $n \ge 6$) and we can use them to choose a $(2,2)$-cycle α_1 supported by $\mathrm{fix}(\sigma)$ and then $\sigma\alpha_1 \in [2, 2, 2]$ and we are done. This completes the proof of Step I in the case $p = 2$.

For the remainder of the proof $p \geq 4$. In particular $p+q \geq 6$. Assume first that σ is a transposition. We may assume that $\mathrm{supp}(\sigma) = \{n-1, n\}$ and notice that $n-2 \geq p+q$ by the assumption. Choose a $(2,2)$-cycle $\alpha_0 \in S_{n-2}$ arbitrarily. Use O_3(c) $\frac{p-2}{2}$ times to find 3-cycles $\beta_1, \ldots, \beta_{(p-2)/2} \in S_{n-2}$ such that $\theta = \alpha_0\beta_1 \cdots \beta_{(p-2)/2}$ is a $(p,2)$-cycle. Use O_3(c) $\frac{q-2}{2}$ times to find 3-cycles $\gamma_1, \ldots, \gamma_{(q-2)/2} \in S_{n-2}$ such that $\theta\gamma_1, \ldots, \gamma_{(q-2)/2}$ is a (p,q)-cycle. Then

$$\sigma\alpha_0\beta_1 \cdots \beta_{(p-2)/2}\gamma_1 \cdots \gamma_{(q-2)/2} \in [p, q, 2]$$

and we are done since $\frac{p-2}{2} + \frac{q-2}{2} + 1 = \frac{p+q-2}{2}$.

Therefore for the remainder of the proof of Step I we assume that σ is not a transposition. Hence, if $r \geq 2$ then $k_1 + k_2 \geq 4$.

Use Lemma 4.12 with $m = p+q+1$ to find 3-cycles $\alpha_1, \ldots, \alpha_\ell$, where $\ell \geq 0$ and $r \geq 2\ell+1$ such that if we set $\tilde{k} = \sum_{i=1}^{2\ell+1} k_i$ and $\xi = \sigma\alpha_1 \cdots \alpha_\ell$ then either

(i) $\tilde{k} \leq p+q$ and $r = 2\ell+2$ and $\xi \in [\tilde{k}, k_{2\ell+2}]$, or

(ii) $\tilde{k} \geq p+q+1$ and $\sum_{i=1}^{2\ell-1} k_i \leq p+q$ and $\xi \in [\tilde{k}, k_{2\ell+2}, \ldots, k_r]$.

Case (i): Observe that $k_{2\ell+2} = n - \tilde{k} \geq n - (p+q) \geq p \geq 4$. Therefore $k_i \geq 4$ for all i and therefore

$$p+q \geq \sum_{i=1}^{2\ell+1} k_i \geq 4(2\ell+1).$$

It follows that $\ell \leq \lfloor \frac{p+q-4}{8} \rfloor$.

Since $\xi \in [\tilde{k}, n-\tilde{k}]$, use O_3(b) to find a 3-cycle $\alpha_{\ell+1}$ such that $\xi\alpha_{\ell+1} \in [n-1, 1]$. Since $n-1 > p+q$, use O_3(a) to find a 3-cycle $\alpha_{\ell+2}$ such that $\xi\alpha_{\ell+1}\alpha_{\ell+2} \in [p, q, n-1-p-q]$ contains a (p,q)-cycle. We are done because $\ell+2 \leq \lfloor \frac{p+q+12}{8} \rfloor$ and one checks that $\lfloor \frac{p+q+12}{8} \rfloor \leq \frac{p+q-2}{2}$ if $p+q \geq 6$.

Case (ii): If $\ell = 0$ then $\sigma = [k_1, \ldots]$, where $k_1 \geq p+q+1$. Then use O_3(a) to find a 3-cycle α_1 such that $\sigma\alpha_1 \in [p, q, k_1 - p - q, \ldots]$ and we are done $\big($since $1 \leq \frac{p+q-2}{2}\big)$. So we only need to consider $\ell \geq 1$.

Suppose first that $\sum_{i=1}^{2\ell-1} k_i = p+q$. Then $\sigma\alpha_1 \cdots \alpha_{\ell-1} \in [p+q, k_{2\ell}, k_{2\ell+1}, \ldots]$. Use O_2(c2) to replace α_ℓ with a $(2,2)$-cycle such that $\sigma\alpha_1 \cdots \alpha_\ell \in [p, q, k_{2\ell} + k_{2\ell+1}, \ldots]$. If $\ell = 1$ then $\ell \leq \frac{p+q-2}{2}$ and we are done. If $\ell \geq 2$ then

$$p+q = \sum_{i=1}^{2\ell-1} k_i = k_1 + k_2 + \sum_{i=3}^{2\ell-1} k_i \geq 4 + 2\ell - 3$$

since $k_i \geq 1$. Therefore $\ell \leq \lfloor \frac{p+q-1}{2} \rfloor = \frac{p+q-2}{2}$ since $p+q$ is even, and we are done.

It remains to consider the case $\sum_{i=1}^{2\ell-1} k_i \leq p+q-1$. Assume first that $k_{2\ell} = 1$. This implies that $k_{2\ell+1} = 1$ and since $\sum_{i=1}^{2\ell+1} k_i \geq p+q+1$ it follows that $\sum_{i=1}^{2\ell} k_i =$

$p+q$, and therefore $\sigma\alpha_1\cdots\alpha_{\ell-1}\in[p+q-1,1,1,\dots]$. Use O_3(b) to replace α_ℓ with a 3-cycle such that $\sigma\alpha_1\cdots\alpha_\ell\in[p,q,1,\dots]$. Since $k_1\geq 2$ and $k_i\geq 1$ we get

$$p+q=\sum_{i=1}^{2\ell}k_i\geq 2+2\ell-1$$

and therefore $\ell\leq\lfloor\frac{p+q-1}{2}\rfloor=\frac{p+q-2}{2}$ and we are done.

Assume that $k_{2\ell}\geq 2$. Since $\tilde{k}\geq p+q+1$, use O_3(a) to find a 3-cycle $\alpha_{\ell+1}$ such that $\xi\alpha_{\ell+1}\in[p,q,\tilde{k}-p-q,\dots]$ contains a (p,q)-cycle. Since $k_1\geq\cdots\geq k_{2\ell}\geq 2$ and $\sum_{i=1}^{2\ell-1}k_i\leq p+q-1$ we deduce that $2(2\ell-1)\leq p+q-1$; hence $\ell\leq\lfloor\frac{p+q+1}{4}\rfloor=\frac{p+q}{4}$. Therefore $\ell+1\leq\lfloor\frac{p+q+4}{4}\rfloor$ and we are done since $\lfloor\frac{p+q+4}{4}\rfloor\leq\frac{p+q}{2}$ if $p+q\geq 6$. This completes the proof of Step I. □

Step II: Let $\mu\in S_n$ contain a (p,q)-cycle. Then $\|\mu\|_\tau\leq\Delta(S_{n-p-q})+2$.

Proof of Step II. We first consider the case $p=2$. Hence $q=2$. Consider $\mu\in S_n$, which contains a $(2,2)$-cycle. We write μ as a product of disjoint permutations $\mu=\mu_0\mu_{2,2}$, where $\mu_{2,2}$ is a $(2,2)$-cycle in S_4 and $\mu_0\in S_{n-4}$. Notice that $n-4=n-(p+q)\geq p=2$. Similarly we write $\tau=\tau_0\tau_{2,2}$. Since τ is an odd permutation and $\tau_{2,2}$ is even, $\tau_0\in S_{n-4}$ is an odd permutation, and by Lemma 4.4 it normally generates it. By Lemma 4.2 there are $\lambda_1,\dots,\lambda_\ell\in S_{n-4}$, where $\ell\leq\Delta(S_{n-4})$ such that

$$\mu_0=\tau_0^{\lambda_1}\cdots\tau_0^{\lambda_\ell}.$$

Suppose that ℓ is odd. Since $|\tau_{2,2}|=2$,

$$\tau^{\lambda_1}\cdots\tau^{\lambda_\ell}=(\tau_0^{\lambda_1}\cdots\tau_0^{\lambda_\ell})\cdot(\tau_{2,2})^\ell=\mu_0\tau_{2,2}$$

is conjugate to μ since both $\mu_{2,2}$ and $\tau_{2,2}$ are $(2,2)$-cycles, so $\|\mu\|_\tau\leq\ell\leq\Delta(S_{n-4})$.

Suppose that ℓ is even. If $\ell=0$ then $\mu=\mu_{2,2}$ is $(2,2)$-cycle. Since τ contains a $(2,2)$-cycle, $\|\mu\|_\tau\leq 2\leq\Delta(S_{n-4})+2$ and we are done. Otherwise $\ell\geq 2$. In this case we choose $\pi\in S_4$ such that $\tau_{2,2}^\pi\tau_{2,2}$ is a $(2,2)$-cycle. Then

$$\tau^{\lambda_1\pi}\tau^{\lambda_2}\cdots\tau^{\lambda_\ell}=(\tau_0^{\lambda_1}\cdots\tau_0^{\lambda_\ell})\cdot(\tau_{2,2}^\pi\tau_{2,2}\cdot(\tau_{2,2})^{\ell-2})=\mu_0\cdot(\tau_{2,2}^\pi\tau_{2,2})$$

is conjugate to μ; hence $\|\mu\|_\tau\leq\ell\leq\Delta(S_{n-4})$ and this completes the proof of Step II in the case $p=2$.

For the remainder of the proof of Step II assume $p\geq 4$. Write $\mu=\mu_0\mu_{p,q}$, a product of disjoint permutations with $\mu_{p,q}\in S_{p+q}$ a (p,q)-cycle and $\mu_0\in S_{n-p-q}$. Notice that $n-p-q\geq p\geq 4$ by assumption. Similarly write $\tau=\tau_0\tau_{p,q}$. Since τ is odd and $\tau_{p,q}$ is even, τ_0 is odd and therefore normally generates S_{n-p-q}. By Lemma 4.2 there are $\lambda_1,\dots,\lambda_\ell\in S_{n-p-q}$, where $\ell\leq\Delta(S_{n-p-q})$, such that

$$\mu_0=\tau_0^{\lambda_1}\cdots\tau_0^{\lambda_\ell}.$$

Choose $\theta\in S_{p+q}$ such that $\tau_{p,q}^\theta=\tau_{p,q}^{-1}$.

If ℓ is odd then

$$\tau^{\lambda_1}\tau^{\lambda_2\theta}\tau^{\lambda_3}\tau^{\lambda_4\theta}\cdots\tau^{\lambda_{\ell-1}\theta}\tau^{\lambda_\ell} = (\tau_0^{\lambda_1}\cdots\tau_0^{\lambda_\ell})\cdot((\tau_{p,q}\tau_{p,q}^{-1})^{(\ell-1)/2}\cdot\tau_{p,q}) = \mu_0\tau_{p,q}$$

is conjugate to μ so $\|\mu\|_\tau \le \ell \le \Delta(S_{n-p-q})$.

Suppose that ℓ is even. Since both p, q are even, $\tau_{p,q}^2$ is a $\left(\frac{p}{2}, \frac{p}{2}, \frac{q}{2}, \frac{q}{2}\right)$-cycle. Use O_2(d) to find a $(2, 2)$-cycle β such that $\tau_{p,q}\beta$ is a (p, q)-cycle.

If $\ell = 0$ then $\mu = \mu_{p,q}$. Choose $\pi \in S_{n-p-q}$ such that $\tau_0^\pi = \tau_0^{-1}$. Then

$$\tau\tau^\pi\beta = (\tau_0\tau_0^{-1})(\tau_{p,q}^2)\beta \in [p, q]$$

is conjugate to μ. Since $p \ge 4$, Lemma 4.7 gives $\|\beta\|_\tau \le 2$ and therefore $\|\mu\| \le \|\tau\|_\tau + \|\tau^\pi\|_\tau + \|\beta\|_\tau \le 4$. By Corollary 4.6 and since $n - p - q \ge p \ge 4$, we get $\Delta(S_{n-p-q}) + 2 \ge 3 + 2 > \|\mu\|_\tau$.

If $\ell \ge 2$ is even then

$$\begin{aligned}\tau^{\lambda_1}\tau^{\lambda_2}\tau^{\lambda_3\theta}\tau^{\lambda_4}\tau^{\lambda_5\theta}\tau^{\lambda_6}\cdots\tau^{\lambda_{\ell-1}\theta}\tau^{\lambda_\ell}\cdot\beta &= (\tau_0^{\lambda_1}\cdots\tau_0^{\lambda_\ell})\cdot(\tau_{p,q}^2(\tau_{p,q}^{-1}\tau_{p,q})^{(\ell-2)/2})\cdot\beta \\ &= \mu_0\cdot\tau_{p,q}^2\cdot\beta\end{aligned}$$

is conjugate to μ since $\tau_{p,q}^2\beta$ is a (p, q)-cycle. Therefore

$$\|\mu\|_\tau \le \ell + \|\beta\|_\tau = \ell + 2 \le \Delta(S_{n-p-q}) + 2.$$

This completes the proof of Step II. □

Step III: We prove that $\|\sigma\|_\tau \le \Delta(S_{n-p-q}) + p + q$.

Proof of Step III. First, consider the case $p = 2$. Hence $q = 2$. By Step I there are $(2, 2)$-cycles $\alpha_1, \ldots, \alpha_t$ where $t \le \frac{p+q-2}{2}$ such that $\mu = \sigma\alpha_1\cdots\alpha_t$ contains a (p, q)-cycle. By Step II, $\|\mu\|_\tau \le \Delta(S_{n-p-q}) + 2$. By Lemma 4.7, $B_\tau(2)$ contains all $(2, 2)$-cycles. Therefore

$$\|\sigma\|_\tau \le \|\mu\|_\tau + 2t \le \Delta(S_{n-p-q}) + 2 + (p + q - 2) = \Delta(S_{n-p-q}) + p + q.$$

If $p \ge 4$ then Lemma 4.7 implies that $B_\tau(2)$ contains all 3-cycles and all $(2, 2)$-cycles. By Step I there are $\alpha_1, \ldots, \alpha_t$ such that $t \le \frac{p+q-2}{2}$ and α_i are either 3-cycles or $(2, 2)$-cycles and $\mu = \sigma\alpha_1\cdots\alpha_t$ contains a (p, q)-cycle. By Step II $\|\mu\|_\tau \le \Delta(S_{n-p-q}) + 2$ so

$$\|\sigma\|_\tau \le \|\mu\|_\tau + 2t \le \Delta(S_{n-p-q}) + 2 + (p + q - 2) = \Delta(S_{n-p-q}) + p + q. \quad \square$$

Proposition 4.15. *Let $\tau \in S_n$ be an n-cycle, $n \ge 4$ even. Then $\|S_n\|_\tau \le n - 1$.*

Proof. First, τ is an odd permutation, and hence normally generates S_n by Lemma 4.4. Since $n \ge 4$, Lemma 4.7 shows that $B_\tau(2)$ contains all 3-cycles and all $(2, 2)$-cycles. Consider some $1 \ne \sigma \in S_n$ with cycle structure $[k_1, \ldots, k_r]$, where $\sum_i k_i = n$ and $k_1 \ge \cdots \ge k_r$. Then $k_1 \ge 2$ since $\sigma \ne 1$. We need to show that $\|\sigma\|_\tau \le n - 1$.

Suppose first that r is odd. If $r = 1$ then σ is an n-cycle, $\|\sigma\|_\tau = 1 \le n-1$ and we are done. If $r \ge 3$, use O_3(a) to find a 3-cycle α_1 such that $\sigma\alpha_1 \in [k_1, k_2, k_3, n-(k_1+k_2+k_3)]$. Repeat this process to find 3-cycles $\alpha_2, \dots, \alpha_{(r-1)/2}$ such that $\sigma\alpha_1 \cdots \alpha_{(r-1)/2} \in [k_1, \dots, k_r]$ (this is possible since r is odd). This shows that

$$\|\sigma\|_\tau \le \tfrac{r-1}{2} \cdot \|\alpha_i\|_\tau \le 2 \cdot \tfrac{r-1}{2} = r-1 \le n-1.$$

Suppose that r is even ($r \ge 2$). Then σ is not a transposition (because in that case r is odd). If σ is either a 3-cycle or a $(2,2)$-cycle then $\|\sigma\|_\tau \le 2$ by Lemma 4.7 and we are done since $n \ge 4$. Therefore either

- $k_1 \ge 4$, in which case $r \le 1 + (n-k_1) \le n-3$, or
- $k_1 = 3$ and $k_2 \ge 2$ in which case $r \le 1+1+(n-5) = n-3$, or
- $k_1 = 2$ and $k_2, k_3 = 2$ (since σ is not a transposition nor a $(2,2)$-cycle), so $r \le 3 + (n-6) = n-3$.

So we may assume that $r \le n-3$.

Since n is even, τ^2 is an $\left(\frac{n}{2}, \frac{n}{2}\right)$-cycle. Since $k_1 \ge \cdots \ge k_r$ and $r \ge 2$ and $\sum_i k_i = n$, we see that $k_r \le \frac{n}{2}$. If $k_r = \frac{n}{2}$ then $r = 2$ and σ is an $\left(\frac{n}{2}, \frac{n}{2}\right)$-cycle, so

$$\|\sigma\|_\tau = \|\tau^2\|_\tau = 2 \le n-1$$

and we are done. So assume $k_r < \frac{n}{2}$. Apply O_3(b) to τ^2 to find a 3-cycle α_0 such that $\tau^2\alpha_0 \in [n-k_r, k_r]$. Apply O_3(a) $\frac{r-2}{2}$ times to find 3-cycles $\alpha_1, \dots, \alpha_t$, where $t = \frac{r-2}{2}$, that split the $(n-k_r)$-cycle into $r-1$ cycles and get $\sigma\alpha_0 \cdots \alpha_t \in [k_1, \dots, k_r]$. Since $\|\alpha_i\|_\tau \le 2$ we get

$$\|\sigma\|_\tau \le 2(t+1) = r \le n-1$$

(because $\sigma \ne 1$). □

Proposition 4.16. *Consider an odd permutation $\tau \in S_n$ and assume that τ fixes a point. Then $\|S_n\|_\tau \le \|S_{n-1}\|_\tau + 1$. In particular $\|S_n\|_\tau \le \Delta(S_{n-1}) + 1$.*

Proof. Up to conjugation we may assume that τ fixes n. Any $\sigma \in S_n$ either fixes a point, in which case up to conjugacy $\sigma \in S_{n-1}$, or there exists τ' conjugate to τ such that $\sigma\tau'$ fixes a point. So up to conjugation $\sigma\tau' \in S_{n-1}$ for some $\tau' \in B_\tau(1)$. Therefore

$$\|\sigma\|_\tau \le \|\sigma\tau'\|_\tau + \|\tau'\|_\tau \le \|S_{n-1}\|_\tau + 1 \le \Delta(S_{n-1}) + 1.$$ □

Proof of Theorem 1.2. We use induction on $n \ge 2$. First, $\Delta(S_2) = 1$ is a triviality and $\Delta(S_3) = 2$ by Theorem 1.1.

Assume inductively that $\Delta(S_m) = m-1$ for all $2 \le m < n$. By Corollary 4.6, $\Delta(S_n) \ge n-1$. To prove equality we need to show that $\|S_n\|_X \le n-1$ for any normally generating set X. By Lemma 4.4, X contains an odd permutation τ which normally generates, and hence $\|S_n\|_X \le \|S_n\|_\tau$. So it suffices to prove that $\|S_n\|_\tau \le n-1$ for any odd permutation τ.

If τ has a fixed point then by Proposition 4.16

$$\|S_n\|_\tau \leq \Delta(S_{n-1}) + 1 \leq n - 2 + 1 = n - 1$$

and we are done. So in order to establish the induction step we need to check that $\|S_n\|_\tau \leq n - 1$ for odd τ without fixed points. Recall Notation 4.11.

For $n = 4$ the only fixed-point free odd permutations are the 4-cycles. If τ is one then $\|S_4\|_\tau \leq 3$ by Proposition 4.15. So $\Delta(S_4) = 3$.

For $n = 5$ the only fixed-point free odd permutations are the $(3, 2)$-cycles. Let τ be one. Then $[3, 2] \subseteq B_\tau(1)$ by definition and $[3] \subseteq B_\tau(2)$ by Lemma 4.7. We apply Proposition 2.2(iii) and O_3(a) to deduce that

$$[2] = [1, 1, 1, 2] \subseteq [3, 2] \cdot [3] \subseteq B_\tau(3)$$

and O_3(b) to deduce that

$$[4] = [1, 4] \subseteq [3, 2] \cdot [3] \subseteq B_\tau(3).$$

Apply O_3(b) to get

$$[2, 2] \subseteq [3, 1] \cdot [3] \subseteq B_\tau(4)$$

and O_3(c) to get

$$[5] \subseteq [3, 1, 1] \cdot [3] \subseteq B_\tau(4).$$

We have exhausted all the nontrivial conjugacy classes in S_5 and therefore $\|S_5\|_\tau \leq 4$ as needed.

For $n = 6$ the only fixed-point free odd permutations are the $(2, 2, 2)$-cycles and 6-cycles. If τ is a 6-cycle then $\|S_6\|_\tau \leq 5$ by Proposition 4.15. Consider $\tau \in [2, 2, 2]$. Then $[2, 2, 2] \subseteq B_\tau(1)$ by definition and $[2, 2] \subseteq B_\tau(2)$ by Lemma 4.7. Now, $[2], [6], [4] \subseteq B_\tau(3)$ because

$$\begin{aligned} [2] = [1, 1, 1, 1, 2] &\subseteq [2, 2, 2] \cdot [2, 2] && \text{by } O_2\text{(b1)}, \\ [6] &\subseteq [2, 2, 2] \cdot [2, 2] && \text{by } O_2\text{(c1)}, \\ [4] = [1, 1, 4] &\subseteq [2, 2, 2] \cdot [2, 2] && \text{by } O_2\text{(c2)}. \end{aligned}$$

Next, $[5], [3], [4, 2], [3, 3] \subseteq B_\tau(4)$ because

$$\begin{aligned} [5] &\subseteq [2, 2, 1] \cdot [2, 2] && \text{by } O_2\text{(c1)}, \\ [3] = [1, 1, 3] &\subseteq [2, 2, 1] \cdot [2, 2] && \text{by } O_2\text{(c2)}, \\ [4, 2] &\subseteq [2, 2, 1, 1] \cdot [2, 2] && \text{by } O_2\text{(d)}, \\ [3, 3] &\subseteq [2, 1, 2, 1] \cdot [2, 2] && \text{by } O_2\text{(d)}. \end{aligned}$$

Finally

$$[3, 2] = [3, 1, 2] \subseteq [6] \cdot [2, 2] \subseteq B_\tau(3 + 2)$$

by $O_2(a)$. This exhausts all the nontrivial conjugacy classes in S_6 and therefore $\|S_6\|_\tau \leq 5$ as needed.

We now assume that $n \geq 7$ and that $\Delta(S_m) = m-1$ for all $2 \leq m < n$. Choose an odd permutation $\tau \in S_n$ without fixed points. If τ is an n-cycle then $\|S_n\|_\tau \leq n-1$ by Proposition 4.15. So we assume that τ is a product of at least two cycles each of length $k \geq 2$. If one of these cycles has odd length $k \geq 3$ then $n-k \geq 2$ (or else τ has a fixed point) and Proposition 4.13, together with the induction hypothesis, shows that

$$\|S_n\|_\tau \leq \Delta(S_{n-k}) + k = n-k-1+k = n-1$$

as needed. If τ contains no cycles of odd length then it is a product of cycles of even length. Since τ is odd, the number of these cycles must be odd, and since τ is not a cycle, it is a product of at least three cycles of even length. Let $p \geq q$ be the lengths of the shortest two cycles in τ. Then $q \geq 2$ and $n-(p+q) \geq p$ because τ contains a third cycle of length at least p. Appealing to Proposition 4.14 and the induction hypothesis, we deduce that

$$\|S_n\|_\tau \leq \Delta(S_{n-p-q}) + p + q = n-1.$$

The induction step is complete. □

References

[Babai and Seress 1992] L. Babai and A. Seress, "On the diameter of permutation groups", *European J. Combin.* **13**:4 (1992), 231–243. MR Zbl

[Babai et al. 1990] L. Babai, G. Hetyei, W. M. Kantor, A. Lubotzky, and A. Seress, "On the diameter of finite groups", pp. 857–865 in *31st Annual Symposium on Foundations of Computer Science, Vol. II* (St. Louis, MO, 1990), IEEE Comput. Soc. Press, Los Alamitos, CA, 1990. MR

[Bardakov et al. 2012] V. Bardakov, V. Tolstykh, and V. Vershinin, "Generating groups by conjugation-invariant sets", *J. Algebra Appl.* **11**:4 (2012), art. id. 1250071. MR Zbl

[Burago et al. 2008] D. Burago, S. Ivanov, and L. Polterovich, "Conjugation-invariant norms on groups of geometric origin", pp. 221–250 in *Groups of diffeomorphisms*, edited by R. Penner et al., Adv. Stud. Pure Math. **52**, Math. Soc. Japan, Tokyo, 2008. MR Zbl

[Hall 1959] M. Hall, Jr., *The theory of groups*, The Macmillan Co., New York, 1959. MR Zbl

[Helfgott and Seress 2014] H. A. Helfgott and A. Seress, "On the diameter of permutation groups", *Ann. of Math.* (2) **179**:2 (2014), 611–658. MR Zbl

[Kędra et al. 2018] J. Kędra, A. Libman, and B. Martin, "Strong and uniform boundedness of groups", preprint, 2018. arXiv

[Klopsch and Lev 2003] B. Klopsch and V. F. Lev, "How long does it take to generate a group?", *J. Algebra* **261**:1 (2003), 145–171. MR Zbl

[Preparata and Vuillemin 1981] F. P. Preparata and J. Vuillemin, "The cube-connected cycles: a versatile network for parallel computation", *Comm. ACM* **24**:5 (1981), 300–309. MR

[Rotman 1973] J. J. Rotman, *The theory of groups: an introduction*, 2nd ed., Allyn and Bacon, Boston, MA, 1973. MR Zbl

[Tarry 2020] C. Tarry, "Strong and uniform boundedness of groups", undergraduate project, 2020.

Received: 2020-03-04 Revised: 2020-05-10 Accepted: 2020-06-25

a.libman@abdn.ac.uk *Institute of Mathematics, University of Aberdeen, King's College, Aberdeen, United Kingdom*

charlotte0687@hotmail.co.uk *Institute of Mathematics, University of Aberdeen, King's College, Aberdeen, United Kingdom*

INVOLVE 13:4 (2020)
https://doi.org/10.2140/involve.2020.13.673

Existence of multiple solutions to a discrete boundary value problem with mixed periodic boundary conditions

Kimberly Howard, Long Wang and Min Wang

(Communicated by Johnny Henderson)

A second-order discrete boundary value problem with mixed periodic boundary conditions is studied. Sufficient conditions on the multiplicity of solutions in a weak sense are obtained by using the critical point theory. An example is given to demonstrate the applications of our results as well.

1. Introduction

We consider a nonlinear boundary value problem (BVP) consisting of a second-order difference equation

$$-\Delta^2 u(t-1) = f(t, u(t)), \quad t \in [2, N]_{\mathbb{Z}}, \tag{1-1}$$

and a pair of mixed periodic boundary conditions (BCs)

$$u(0) = -u(N), \quad \Delta u(0) = \Delta u(N), \tag{1-2}$$

where

- $N \geq 3$ is a positive integer and $[a, b]_{\mathbb{Z}}$ denotes the discrete interval $\{a, a+1, \ldots, b\}$ for any integers a and b with $a \leq b$;
- Δ is the forward difference operator defined by $\Delta u(t) = u(t+1) - u(t)$ and $\Delta^2 u(t) = \Delta(\Delta u(t))$; and
- $f : [2, N]_{\mathbb{Z}} \times \mathbb{R} \to \mathbb{R}$ is continuous in its second variable.

By a solution of BVP (1-1), (1-2), we mean a function $u : [0, N+1]_{\mathbb{Z}} \to \mathbb{R}$ that satisfies (1-1) and (1-2).

Remark 1.1. It is notable that in this paper, the solution is only required to satisfy (1-1) on $[2, N]_{\mathbb{Z}}$ instead of $[1, N]_{\mathbb{Z}}$. We call it *a solution in a weak sense*. Assume (1-1) is defined on $[1, N]_{\mathbb{Z}}$. By applying the standard variational technique in the

MSC2010: 34B15, 39A23.

Keywords: discrete boundary value problem, mixed periodic boundary conditions, variational methods.

traditional way, let

$$\widetilde{H} = \{u : [0, N+1]_{\mathbb{Z}} \to \mathbb{R} \mid u \text{ satisfies the BC (1-2)}\},$$
$$\tilde{J}(u) = \frac{1}{2}\sum_{t=1}^{N}(\Delta u(t-1))^2 - \sum_{t=1}^{N}\int_0^{u(t)} f(t,s)\,ds.$$

If $u \in \widetilde{H}$ is a critical point of $\tilde{J}(u)$, then u must satisfy $\Delta u(0) = \Delta u(N) = 0$ and $u(0) = -u(N)$, which become a special case of the antiperiodic BCs studied in the literature; see for example [Lyons and Neugebauer 2015a; Kuang and Yang 2018]. We are not interested in such simple BCs. This is the motivation to consider the solution in a weak sense. The reader is referred to [Kong and Wang ≥ 2020, Remark 2.2] for more discussions on the solution in a weak sense.

BVPs have been an active research area for decades. Both continuous and discrete BVPs subject to various BCs have been investigated by numerous scholars. The reader is referred to [Davis et al. 2016; Feng et al. 2018; Garcia and Neugebauer 2019; Graef et al. 2008; 2013; 2017; 2018; Henderson and Luca 2018; 2019; Kuang and Yang 2018; Lyons and Neugebauer 2015a; 2015b; Liang and Weng 2007; Liu et al. 2018; Kong et al. 2017] for recent advances in BVPs. Although BVPs with either periodic BCs or antiperiodic BCs have been extensively studied, to the best of our knowledge, there are relatively few works on the problems with mixed periodic BCs, i.e., a combination of periodic and antiperiodic BCs. For continuous BVPs, Garcia and Neugebauer [2019] studied the existence of solutions for the BVPs consisting of the differential equation

$$u'' = f(t, u, u'), \quad 0 < t < 1,$$

and two sets of mixed periodic BCs

$$u(0) + u(1) = 0, \quad u'(0) - u'(1) = 0,$$

or

$$u(0) - u(1) = 0, \quad u'(0) + u'(1) = 0.$$

For discrete BVPs, Kong and Wang [2020] considered the existence of solutions (in a weak sense) for the BVP

$$\begin{aligned} -\Delta^2 u(t-1) &= f(u(t)), \quad t \in [2, N]_{\mathbb{Z}}, \\ u(0) &= -u(N), \quad \Delta u(0) = \Delta u(N), \end{aligned} \tag{1-3}$$

by the mountain pass lemma. In [Kong and Wang ≥ 2020], the existence of solutions (in a weak sense) for the BVP

$$\begin{aligned} -\Delta(r(t-1)\Delta u(t-1)) &= f(t, u(t)), \quad t \in [2, N]_{\mathbb{Z}}, \\ u(0) &= u(N), \quad r(0)\Delta u(0) = -r(N)\Delta u(N), \end{aligned}$$

was studied by Clark's theorem. Note that as commented in [Kong and Wang $\geq$ 2020], there was a typo in (1.1) in [Kong and Wang 2020]. The domain was mistakenly written as $t \in [1, N]_{\mathbb{Z}}$ and should be replaced by $t \in [2, N]_{\mathbb{Z}}$ as seen in (1-3) above.

We will extend [Kong and Wang 2020] by considering a more general problem. It is easy to see that BVP (1-1), (1-2) covers BVP (1-3) as a special case. By using the linking theorem, a new set of sufficient conditions for the existence of solutions will be derived. Our work will supplement the existing results and contribute to the application of the variational method to the BVPs with mixed periodic BCs.

This paper is organized as follows: after this introduction, the preliminary results are given in Section 2. Section 3 contains the main results. An example is given in Section 3 as well.

2. Preliminaries

We present some definitions and lemmas needed for the proof of our main result.

Definition 2.1. Assume U is a Banach space. We say that a functional $J \in C^1(U, \mathbb{R})$ satisfies the Palais–Smale (PS) condition if every sequence $\{u_n\} \subset U$, such that $J(u_n)$ is bounded and $J'(u_n) \to 0$ as $n \to \infty$, has a convergent subsequence. The sequence $\{u_n\}$ is called a PS sequence.

The following lemma, which is the well-known linking theorem, plays a crucial role in the proof of our main results; see for example [Rabinowitz 1986, Theorem 5.3].

Lemma 2.2. *Let U be a real Banach space, $U = U_1 \oplus U_2$, where U_1 is finite-dimensional. Suppose that $J \in C^1(U, \mathbb{R})$ satisfies the PS condition and the following*:

(J1) *There are positive constants c and ρ such that $J|_{\partial B_\rho(0) \cap U_2} \geq c$.*

(J2) *There are $\mu \in \partial B_1(0) \cap U_2$ and a positive constant $\hat{c} \geq \rho$ such that $J|_{\partial\Omega} \leq 0$, where*

$$\Omega = (\bar{B}_{\hat{c}}(0) \cap U_1) \oplus \{s\mu \mid 0 < s < \hat{c}\}.$$

Then J possesses a critical value $c_0 \geq c$, where

$$c_0 = \inf_{d \in \Gamma} \sup_{u \in \Omega} J(d(u)) \tag{2-1}$$

and $\Gamma = \{d \in C(\bar{\Omega}, U) \mid d|_{\partial\Omega} = I\}$, where I denotes the identity operator.

In the rest of the paper, we will use the Banach space U defined by

$$U = \{u : [0, N+1]_{\mathbb{Z}} \to \mathbb{R} \mid u(0) = -u(N),\ \Delta u(0) = \Delta u(N),\ u(1) = 0\}.$$

By [Kong and Wang 2020, Remark 2.1], U is an $(N-1)$-dimensional Banach space with the norm defined by

$$\|u\| = \left(\sum_{t=1}^{N} u^2(t)\right)^{\frac{1}{2}}.$$

Let $\tilde{f} : [1, N]_{\mathbb{Z}} \times \mathbb{R} \to \mathbb{R}$ and $\tilde{F} : [1, N]_{\mathbb{Z}} \times \mathbb{R} \to \mathbb{R}$ be defined by

$$\tilde{f}(t, x) = \begin{cases} 0, & t = 1, \\ f(t, x), & t \in [2, N-1]_{\mathbb{Z}}, \\ f(t, x) + 2x, & t = N, \end{cases} \tag{2-2}$$

$$\tilde{F}(t, x) = \int_0^x \tilde{f}(t, s)\, ds. \tag{2-3}$$

Clearly, both $\tilde{f}(t, x)$ and $\tilde{F}(t, x)$ are continuous in x. Define $J : U \to \mathbb{R}$ by

$$J(u) = \frac{1}{2}\sum_{t=1}^{N}(\Delta u(t-1))^2 - \sum_{t=1}^{N}\tilde{F}(t, u(t)). \tag{2-4}$$

A critical point of J is a point u at which J', the Fréchet derivative of J, vanishes, and the value of J at u is then called a critical value of J; see for example [Rabinowitz 1986]. The following lemma links the critical point of J and the solution of BVP (1-1), (1-2).

Lemma 2.3. *The function $u \in U$ is a critical point of J if and only if u is a solution of BVP* (1-1), (1-2) *in a weak sense.*

Proof. For any $u \in U$, from (2-2)–(2-4), we obtain

$$J(u) = \frac{1}{2}\sum_{t=1}^{N}(\Delta u(t-1))^2 - \sum_{t=2}^{N}\int_0^{u(t)} f(t, s)\, ds - 2\int_0^{u(N)} s\, ds.$$

It is easy to see that J is continuously differentiable. For any $v \in U$, we have that $J'(u)$ at $u \in U$ is given by

$$\langle J'(u), v\rangle = \sum_{t=1}^{N}\Delta u(t-1)\Delta v(t-1) - \sum_{t=2}^{N} f(t, u(t))v(t) - 2u(N)v(N).$$

Using the summation by parts formula, we have

$$\begin{aligned}\sum_{t=1}^{N}\Delta u(t-1)\Delta v(t-1) &= \Delta u(N)v(N) - \Delta u(0)v(0) - \sum_{t=1}^{N}\Delta^2 u(t-1)v(t) \\ &= 2\Delta u(N)v(N) - \sum_{t=1}^{N}\Delta^2 u(t-1)v(t).\end{aligned}$$

Then it follows that

$$\begin{aligned}\langle J'(u), v\rangle &= 2\Delta u(N)v(N) - \sum_{t=1}^{N} \Delta^2 u(t-1)v(t) - \sum_{t=2}^{N} f(t, u(t))v(t) - 2u(N)v(N)\\ &= -2\Delta u(0)v(0) - \sum_{t=1}^{N} \Delta^2 u(t-1)v(t) - \sum_{t=2}^{N} f(t, u(t))v(t) - 2u(0)v(0)\\ &= -\sum_{t=1}^{N} \Delta^2 u(t-1)v(t) - \sum_{t=2}^{N} f(t, u(t))v(t)\\ &= -\sum_{t=2}^{N} [\Delta^2 u(t-1) + f(t, u(t))]v(t).\end{aligned}$$

Therefore $\langle J'(u), v\rangle = 0$ if and only if (1-1) holds. □

Now we consider an equivalent form of J. Let

$$A = \begin{bmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 & 1\\ -1 & 2 & -1 & 0 & \cdots & \cdots & 0\\ 0 & -1 & 2 & -1 & 0 & \cdots & 0\\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ 0 & \cdots & 0 & -1 & 2 & -1 & 0\\ 0 & \cdots & \cdots & 0 & -1 & 2 & -1\\ 1 & 0 & \cdots & 0 & 0 & -1 & 2 \end{bmatrix}_{N\times N}, \quad N \geq 3. \tag{2-5}$$

Then it is easy to verify that for any $u \in U$,

$$J(u) = \tfrac{1}{2} u^T A u - \sum_{t=1}^{N} \tilde{F}(t, u(t)), \tag{2-6}$$

where $(\cdot)^T$ denotes the transpose and $u = (0, u(2), \ldots, u(N))^T$. Note that by [Kong and Wang 2020, Remark 2.2], A has N positive eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$ such that for any $u \in \mathbb{R}^N$, we have $\lambda_1 \|u\|^2 \leq u^T A u \leq \lambda_N \|u\|^2$.

3. Main results

The following theorem is our first main result.

Theorem 3.1. *Let $\lambda_1, \ldots, \lambda_N$ be the eigenvalues of A defined by* (2-5). *Assume the function f from* (1-1) *satisfies $f(t, 0) = 0$, $t \in [2, N]_{\mathbb{Z}}$. Additionally, the function $\tilde{F}(t, x)$ defined by* (2-3) *satisfies the following assumptions*:

(F1) *There exist two constants $\delta_1 > 0$ and $a_1 \in \left(0, \frac{1}{2}\lambda_1\right)$ such that $\tilde{F}(t, x) \leq a_1 x^2$ for $|x| \leq \delta_1$ and $t \in [2, N]_{\mathbb{Z}}$.*

(F2) *There exist two constants* $a_2 \in \left(\frac{1}{2}\lambda_N, \infty\right)$ *and* $a_3 > 0$ *such that* $\tilde{F}(t,x) \geq a_2 x^2 - a_3$ *for all* $(t,x) \in [2,N]_{\mathbb{Z}} \times \mathbb{R}$.

Then BVP (1-1), (1-2) *admits at least three solutions, one being trivial, and two being nontrivial.*

The following lemma will be needed to prove Theorem 3.1.

Lemma 3.2. *Assume* (F2) *of Theorem 3.1 is satisfied. Then the functional* J *defined by* (2-4) (*or* (2-6)) *satisfies the PS condition.*

Proof. For any sequence $\{u_k\} \subseteq U$ with $J'(u_k) \to 0$ as $k \to \infty$ and $-B \leq J(u_k) \leq B$, where B is a positive constant, we claim $\{u_k\}$ is bounded. We will prove it by contradiction.

Assume $\{u_k\} \subseteq U$ is an unbounded sequence. Then there exists a subsequence $\{u_m\}$ of $\{u_k\}$ such that $\|u_m\| \to \infty$ and $J'(u_m) \to 0$ as $m \to \infty$, and $-B \leq J(u_m) \leq B$ for $m = 1, 2, \ldots$. From (F2) of Theorem 3.1 and (2-6), when m is large enough, we have

$$\begin{aligned} J(u_m) &= \tfrac{1}{2}u_m^T A u_m - \sum_{t=1}^{N} \tilde{F}(t, u_m(t)) \\ &\leq \tfrac{1}{2}\lambda_N \|u_m\|^2 - \sum_{t=2}^{N} \tilde{F}(t, u_m(t)) \leq \tfrac{1}{2}\lambda_N \|u_m\|^2 - \sum_{t=2}^{N} (a_2 (u_m(t))^2 - a_3) \\ &\leq \tfrac{1}{2}\lambda_N \|u_m\|^2 - a_2 \|u_m\|^2 + a_3 N = \left(\tfrac{1}{2}\lambda_N - a_2\right)\|u_m\|^2 + a_3 N. \end{aligned} \tag{3-1}$$

Hence $J(u_m) \to -\infty$ as $m \to \infty$, and this contradicts the fact that $\{J(u_m)\}$ is bounded. Thus, $\{u_k\}$ is bounded. Since U is a finite-dimensional space, we know $\{u_k\}$ has a convergent subsequence. Therefore, J satisfies the PS condition. □

Now we are ready to prove Theorem 3.1.

Proof of Theorem 3.1. It is obvious that $u \equiv 0$ is a solution of BVP (1-1), (1-2).

By (3-1), we can show $J(u) \to -\infty$ as $\|u\| \to \infty$ for any $u \in U$. Using the fact that $J(u)$ is bounded above in U, define $\bar{J} = \sup_{u \in U} J(u)$. Therefore, we have a sequence $\{u_k\}$ on U such that $\bar{J} = \lim_{k \to \infty} J(u_k)$. From the proof of Lemma 3.2, it follows that $\{u_k\}$ is bounded. Then $\{u_k\}$ has a convergent subsequence defined by $\{u_{k_n}\}$. Let $\bar{u} = \lim_{n \to \infty} u_{k_n}$, and it is easy to see that $\bar{u} \in U$ with $J(\bar{u}) = \bar{J}$. Therefore $\bar{u}$ is a critical point of J.

To apply Lemma 2.2, we need to separate U into two subspaces P and Q with $U = P \oplus Q$. Let $\{\eta_1, \ldots, \eta_N\} \subset \mathbb{R}^N$ be the orthonormal eigenvectors associated with $\{\lambda_1, \ldots, \lambda_N\}$. Define $P = \operatorname{span}\{\eta_1, \ldots, \eta_M\}$ and $Q = \operatorname{span}\{\eta_{M+1}, \ldots, \eta_N\}$, for some $M \in [2, N-1]_{\mathbb{Z}}$, such that $U = P \oplus Q$, $P \cap U \neq \{0\}$, and $Q \cap U \neq \{0\}$. Clearly, both P and Q are finite-dimensional spaces. Moreover, for any $u \in P$, there

exist $b_1, \dots, b_M \in \mathbb{R}$ such that $u = \sum_{i=1}^{M} b_i \eta_i$ and $\|u\|^2 = \sum_{i=1}^{M} b_i^2$; for any $u \in Q$, there exist $b_{M+1}, \dots, b_N \in \mathbb{R}$ such that $u = \sum_{i=M+1}^{N} b_i \eta_i$ and $\|u\|^2 = \sum_{i=M+1}^{N} b_i^2$.

Since U is finite-dimensional, for any $u \in U$, there exist positive numbers κ_1 and κ_2 such that $\kappa_1 < \kappa_2$ and

$$\kappa_1 \max\{|u(1)|, |u(2)|, \dots, |u(N)|\} \le \|u\| \le \kappa_2 \max\{|u(1)|, |u(2)|, \dots, |u(N)|\}.$$

Observe for any $u \in Q$, such that $\|u\| = \kappa_1 \delta_1$ and $\max\{|u(1)|, |u(2)|, \dots, |u(N)|\} \le \delta_1$,

$$\begin{aligned} J(u) &= \tfrac{1}{2} u^T A u - \sum_{t=1}^{N} \tilde{F}(t, u(t)) \ge \tfrac{1}{2}\lambda_1 \|u\|^2 - \sum_{t=2}^{N} \tilde{F}(t, u(t)) \\ &\ge \tfrac{1}{2}\lambda_1 \|u\|^2 - \sum_{t=2}^{N} a_1 (u(t))^2 = \tfrac{1}{2}\lambda_1 \|u\|^2 - a_1 \|u\|^2 \\ &= \left(\tfrac{1}{2}\lambda_1 - a_1\right)(\kappa_1 \delta_1)^2 > 0. \end{aligned} \tag{3-2}$$

Then assumption (J1) of Lemma 2.2 is satisfied. It is easy to see (3-2) also implies $\bar{J} > 0$. So $\bar{u}$ is a nontrivial solution of BVP (1-1), (1-2).

Next we verify (J2) of Lemma 2.2. Choose $\mu \in Q$ and a positive constant $\hat{c}$ such that $\|\mu\| = 1$ and

$$\hat{c} > \max\left\{\kappa_1 \delta_1, \frac{2a_3 N}{2a_2 - \lambda_N}\right\}.$$

Let $\Omega = (B_{\hat{c}}(0) \cap P) \oplus \{s\mu \mid 0 < s < \hat{c}\}$. Note that for any $\nu \in P$

$$\begin{aligned} J(s\mu + \nu) &= \tfrac{1}{2}(s\mu + \nu)^T A (s\mu + \nu) - \sum_{t=1}^{N} \tilde{F}(t, s\mu(t) + \nu(t)) \\ &\le \tfrac{1}{2}\lambda_N s^2 + \tfrac{1}{2}\lambda_N \|\nu\|^2 - \sum_{t=2}^{N} (a_2 (s\mu(t) + \nu(t))^2 - a_3) \\ &\le \tfrac{1}{2}\lambda_N s^2 + \tfrac{1}{2}\lambda_N \|\nu\|^2 - a_2 s^2 - a_2 \|\nu\|^2 + a_3 N \\ &= \left(\tfrac{1}{2}\lambda_N - a_2\right) s^2 + \left(\tfrac{1}{2}\lambda_N - a_2\right) \|\nu\|^2 + a_3 N \\ &\le \left(\tfrac{1}{2}\lambda_N - a_2\right) \|\nu\|^2 + a_3 N. \end{aligned}$$

Therefore, $J|_{\partial\Omega} \le 0$. Hence (J2) holds.

Then by Lemma 2.2, we know $J(u)$ has a critical value $c_0 \ge \left(\frac{1}{2}\lambda_1 - a_1\right)(\kappa_1 \delta_1)^2$. Let $\tilde{u} \in U$ be a critical point corresponding to c_0; i.e., $J(\tilde{u}) = c_0$. If $\bar{u} \ne \tilde{u}$, then by Lemma 2.3, $\bar{u}$ and $\tilde{u}$ are two different solutions of BVP (1-1), (1-2). If $\bar{u} = \tilde{u}$, by (2-1), $\sup_{u \in U} J(u) = \inf_{d \in \Gamma} \sup_{u \in \Omega} J(d(u))$. Choose distinct $d_1, d_2 \in \Gamma$ such that

$$\{d_1(u) \mid u \in \Omega \setminus \partial\Omega\} \cap \{d_2(u) \mid u \in \Omega \setminus \partial\Omega\} = \varnothing.$$

Then

$$\sup_{u\in\Omega} J(d_1(u)) = \sup_{u\in\Omega} J(d_2(u)) = \sup_{u\in U} J(u).$$

Hence there exist $u_1, u_2 \in \Omega$ such that $d_1(u_1)$ and $d_2(u_2)$ are two different critical points of J.

Therefore, we have shown BVP (1-1), (1-2) admits at least three solutions. □

The following corollary is a direct consequence of Theorem 3.1.

Corollary 3.3. *Assume the function f from* (1-1) *satisfies $f(t,0) = 0,\ t \in [2,N]_{\mathbb{Z}}$, and $\tilde{F}(t,x)$ defined by* (2-3) *is of the form*

$$\tilde{F}(t,x) = a(t)|x|^{r(t)}, \quad a(t) > 0,\ r(t) > 2,\ t \in [2,N]_{\mathbb{Z}}.$$

Then BVP (1-1), (1-2) *admits at least three solutions, one being trivial, and two being nontrivial.*

We conclude the paper with the following example.

Example 3.4. Consider BVP (1-1), (1-2) with $N = 4$ and f defined by

$$f(t,x) = \begin{cases} \frac{2}{5}x^3, & t = 2,3, \\ \frac{2}{5}x^3 - 2x, & t = 4. \end{cases}$$

By (2-2) and (2-3), we know

$$\tilde{F}(t,x) = \tfrac{1}{10}x^4, \quad t = 2,3,4.$$

Then by Corollary 3.3, BVP (1-1), (1-2) admits at least three solutions.

Acknowledgement

This research was supported by the Mentor Protégé Research Program from the College of Science and Mathematics at Kennesaw State University.

References

[Davis et al. 2016] J. M. Davis, P. W. Eloe, J. R. Graef, and J. Henderson, "Positive solutions for a singular fourth order nonlocal boundary value problem", *Int. J. Pure Appl. Math.* **109**:1 (2016), 67–84.

[Feng et al. 2018] Y. Feng, J. R. Graef, L. Kong, and M. Wang, "The forward and inverse problems for a fractional boundary value problem", *Appl. Anal.* **97**:14 (2018), 2474–2484. MR Zbl

[Garcia and Neugebauer 2019] A. E. Garcia and J. T. Neugebauer, "Solutions of boundary value problems at resonance with periodic and antiperiodic boundary conditions", *Involve* **12**:1 (2019), 171–180. MR Zbl

[Graef et al. 2008] J. R. Graef, L. Kong, and H. Wang, "Existence, multiplicity, and dependence on a parameter for a periodic boundary value problem", *J. Differential Equations* **245**:5 (2008), 1185–1197. MR Zbl

[Graef et al. 2013] J. R. Graef, L. Kong, and M. Wang, "Multiple solutions to a periodic boundary value problem for a nonlinear discrete fourth order equation", *Adv. Dyn. Syst. Appl.* **8**:2 (2013), 203–215. MR Zbl

[Graef et al. 2017] J. R. Graef, L. Kong, Q. Kong, and M. Wang, "On a fractional boundary value problem with a perturbation term", *J. Appl. Anal. Comput.* **7**:1 (2017), 57–66. MR

[Graef et al. 2018] J. R. Graef, S. Heidarkhani, L. Kong, and M. Wang, "Existence of solutions to a discrete fourth order boundary value problem", *J. Difference Equ. Appl.* **24**:6 (2018), 849–858. MR Zbl

[Henderson and Luca 2018] J. Henderson and R. Luca, "Positive solutions for a system of coupled fractional boundary value problems", *Lith. Math. J.* **58**:1 (2018), 15–32. MR Zbl

[Henderson and Luca 2019] J. Henderson and R. Luca, "Existence of positive solutions for a system of semipositone coupled discrete boundary value problems", *J. Difference Equ. Appl.* **25**:4 (2019), 516–541. MR Zbl

[Kong and Wang 2020] L. Kong and M. Wang, "Existence of solutions for a second order discrete boundary value problem with mixed periodic boundary conditions", *Appl. Math. Lett.* **102** (2020), art. id. 106138. MR Zbl

[Kong and Wang ≥ 2020] L. Kong and M. Wang, "Multiple and particular solutions of a second order discrete boundary value problem with mixed periodic boundary conditions", submitted.

[Kong et al. 2017] L. Kong, J. Parsley, K. Rizzo, and N. Russell, "Anti-periodic solutions for a higher order difference equation with p-Laplacian", *J. Appl. Anal.* **23**:2 (2017), 111–125. MR Zbl

[Kuang and Yang 2018] J. Kuang and Y. Yang, "Variational approach to anti-periodic boundary value problems involving the discrete p-Laplacian", *Bound. Value Probl.* **2018** (2018), art. id. 86. MR

[Liang and Weng 2007] H. Liang and P. Weng, "Existence and multiple solutions for a second-order difference boundary value problem via critical point theory", *J. Math. Anal. Appl.* **326**:1 (2007), 511–520. MR Zbl

[Liu et al. 2018] X. Liu, T. Zhou, and H. Shi, "Existence and multiple solutions to a discrete fourth order boundary value problem", *Adv. Difference Equ.* **2018** (2018), art. id. 427. MR Zbl

[Lyons and Neugebauer 2015a] J. W. Lyons and J. T. Neugebauer, "A difference equation with anti-periodic boundary conditions", *Dyn. Contin. Discrete Impuls. Syst. Ser. A Math. Anal.* **22**:1 (2015), 47–60. MR Zbl

[Lyons and Neugebauer 2015b] J. W. Lyons and J. T. Neugebauer, "Existence of an antisymmetric solution of a boundary value problem with antiperiodic boundary conditions", *Electron. J. Qual. Theory Differ. Equ.* **2015** (2015), art. id. 72. MR Zbl

[Rabinowitz 1986] P. H. Rabinowitz, *Minimax methods in critical point theory with applications to differential equations*, CBMS Region. Conf. Ser. Math. **65**, Amer. Math. Sci., Providence, RI, 1986. MR Zbl

Received: 2020-03-09 Revised: 2020-03-27 Accepted: 2020-04-28

khowar43@students.kennesaw.edu *Department of Mathematics, Kennesaw State University, Marietta, GA, United States*

lwang17@kennesaw.edu *Department of Mathematics, Kennesaw State University, Marietta, GA, United States*

min.wang@kennesaw.edu *Department of Mathematics, Kennesaw State University, Marietta, GA, United States*

INVOLVE 13:4 (2020)
https://doi.org/10.2140/involve.2020.13.683

Minimal flag triangulations of lower-dimensional manifolds

Christin Bibby, Andrew Odesky, Mengmeng Wang, Shuyang Wang, Ziyi Zhang and Hailun Zheng

(Communicated by Kenneth S. Berenhaut)

We prove the following results on flag triangulations of 2- and 3-manifolds. In dimension 2, we prove that the vertex-minimal flag triangulations of $\mathbb{R}P^2$ and $\mathbb{S}^1 \times \mathbb{S}^1$ have 11 and 12 vertices, respectively. In general, we show that $8+3k$ (resp. $8+4k$) vertices suffice to obtain a flag triangulation of the connected sum of k copies of $\mathbb{R}P^2$ (resp. $\mathbb{S}^1 \times \mathbb{S}^1$). In dimension 3, we describe an algorithm based on the Lutz–Nevo theorem which provides supporting computational evidence for the following generalization of the Charney–Davis conjecture: for any flag 3-manifold, $\gamma_2 := f_1 - 5f_0 + 16 \geq 16\beta_1$, where f_i is the number of i-dimensional faces and β_1 is the first Betti number over a field $\boldsymbol{k}$. The conjecture is tight in the sense that for any value of β_1, there exists a flag 3-manifold for which the equality holds.

1. Introduction

A simplicial complex is called flag if all of its minimal nonfaces have cardinality 2. The notion of flagness arises naturally from differential geometry. Gromov [1987] noticed that when a piecewise Euclidean cubical complex is associated with the right-angled metric, the property that the complex is nonpositively curved is equivalent to the condition that every vertex link in the complex is flag. Many classes of simplicial complexes with interesting combinatorial structures are flag, for example, the barycentric subdivisions of simplicial complexes, order complexes of posets, and Coxeter complexes.

In this paper, we focus on a fundamental problem in combinatorial topology of flag complexes: what is the minimum number of vertices that a flag triangulation of a manifold can have? For instance, the boundary complex of a d-simplex gives the minimal triangulation of $\mathbb{S}^{d-1}$. In contrast, a flag triangulation of $\mathbb{S}^{d-1}$ requires at least $2d$ vertices. The vertex-minimal flag triangulation of $\mathbb{S}^{d-1}$ is given by the

MSC2020: 05E45, 57Q15.
Keywords: flag complexes, triangulations of manifolds, Charney–Davis conjecture.

octahedral $(d-1)$-sphere and is unique. However, to the best knowledge of the authors, the vertex-minimal flag triangulations of any other manifolds (even the nonspherical surfaces) remain unknown.

We remark that it is not possible for us to detect flag triangulations of surfaces from existing references: most enumerative results search triangulated manifolds with up to a certain number of vertices, while we expect that the minimal flag triangulations require significantly more vertices than nonflag ones. For example, in [Sulanke and Lutz 2009] it is shown that there are 645592 distinct triangulations of $\mathbb{S}^1 \times \mathbb{S}^1$ with up to 12 vertices, and we find that exactly one of these triangulations is flag. Even though there are enumerations of small triangulations of surfaces, see, e.g., [Sulanke and Lutz 2009; Lutz 2011], in most of the references only the f-vectors and types of manifolds are given, which is not sufficient to check flagness.

The first part of our paper aims at finding the minimal flag triangulations for several surfaces. The method is based on a theorem in [Lutz and Nevo 2016], which states that any two flag homeomorphic PL manifolds can be connected via a sequence of edge subdivisions or admissible edge contractions. By computer search we find

- two nonisomorphic 11-vertex flag triangulations of the real projective plane,
- one 12-vertex flag triangulation of the torus, and
- 28 nonisomorphic 14-vertex flag triangulations of the Klein bottle.

Based on the properties of flagness, we further prove that the 11-vertex and the 12-vertex flag triangulations of $\mathbb{R}P^2$ and $\mathbb{S}^1 \times \mathbb{S}^1$ are indeed vertex-minimal. By defining a new connected sum for flag complexes, we also generate small flag triangulations of all other surfaces. Specifically, we show that $8+4k$ (resp. $8+3k$) vertices suffice to give a flag triangulation of the connected sum of k tori (resp. real projective planes).

The second part of this paper explores γ_2-minimal flag 3-manifolds. The celebrated Charney–Davis conjecture [1995] asserts that for any flag $(2n-1)$-dimensional simplicial sphere Δ,

$$(-1)^n\big(1-\tfrac{1}{2}f_0(\Delta)+\tfrac{1}{4}f_1(\Delta)-\cdots+\big(-\tfrac{1}{2}\big)^{2n}f_{2n-1}(\Delta)\big)\geq 0,$$

where f_i is the number of i-dimensional faces. The Charney–Davis conjecture is an implication of the long-standing Hopf–Chern–Thurston conjecture from a combinatorial perspective; see [Charney and Davis 1995; Forman 2007] for motivation of the conjecture. The conjecture is proved for $n=2$ in [Davis and Okun 2001] using heavy machinery in differential geometry and is now known as the Davis–Okun theorem. In this case, the above inequality can be rephrased using γ-numbers as $\gamma_2(\Delta) := f_1(\Delta)-5f_0(\Delta)+16\geq 0$. (We refer to [Gal 2005] for background on the γ-numbers and other lower-bound-type conjectures for flag spheres.) It is natural to ask if the Davis–Okun theorem has an extension for flag 3-manifolds. In this

paper we explore the connection between the minimal γ_2 for flag triangulations of a 3-manifold M and the first Betti number of M. Based on all flag 3-manifolds that we've constructed, we propose a conjecture that $\gamma_2 \geq 16\beta_1$. Furthermore, we prove that for any integer $b \geq 0$, there exists a flag 3-manifold that attains $\gamma_2 = 16b$. See Section 5 for more discussion.

The structure of the paper is as follows. In Section 2, we review the basic definitions and properties of flag complexes. In Section 3, we present the algorithm for computer search and describe the small triangulations of several surfaces that we have found. In Section 4, we prove that the minimal flag triangulations of $\mathbb{R}P^2$ and $\mathbb{S}^1 \times \mathbb{S}^1$ have exactly 11 and 12 vertices, respectively, and describe how to construct small flag triangulations of other surfaces. We close in Section 5 by discussing potential extension to a manifold Charney–Davis conjecture based on computer search results on flag 3-manifolds.

2. Preliminaries

A *simplicial complex* Δ on a vertex set $V = V(\Delta)$ is a collection of subsets $\sigma \subseteq V$, called *faces*, that is closed under inclusion. Two examples of simplicial complexes are the d-dimensional simplex on V, $\overline{V} := \{\tau : \tau \subseteq V\}$, and its boundary complex, $\partial\overline{V} := \{\tau : \tau \subsetneq V\}$. For $\sigma \in \Delta$, let $\dim \sigma := |\sigma| - 1$ and define the *dimension* of Δ, $\dim \Delta$, as the maximal dimension of its faces. A maximal under inclusion face of Δ is called a *facet*. If all facets of Δ are of the same dimension, then Δ is called *pure*. Let $E(\Delta)$ be the set of 1-faces in Δ. For brevity, we write $\{v\}$ as v.

For a $(d-1)$-dimensional simplicial complex Δ, we let $f_i = f_i(\Delta)$ be the number of i-dimensional faces of Δ for $-1 \leq i \leq d-1$. The vector $(f_{-1}, f_0, \ldots, f_{d-1})$ is called the *f-vector* of Δ. If Δ is a simplicial complex and σ is a face of Δ, the *link* of σ in Δ is $\mathrm{lk}(\sigma, \Delta) := \{\tau - \sigma \in \Delta : \sigma \subseteq \tau \in \Delta\}$, and the *star* of σ in Δ is $\mathrm{st}(\sigma, \Delta) := \{\tau \in \Delta : \sigma \cup \tau \in \Delta\}$. When the context is clear, we will abbreviate the notation and write them as $\mathrm{lk}(\sigma)$ and $\mathrm{st}(\sigma)$ respectively. The *restriction* of Δ to a vertex set W is defined as $\Delta[W] := \{\sigma \in \Delta : \sigma \subseteq W\}$.

If Δ and Γ are simplicial complexes defined on disjoint vertex sets, we define the *join* of Δ and Γ to be the simplicial complex $\Delta * \Gamma = \{\sigma \cup \tau : \sigma \in \Delta, \tau \in \Gamma\}$. Given a face $\sigma \in \Delta$, the *stellar subdivision of Δ along σ* is

$$\mathrm{sd}(\sigma, \Delta) = \{\tau \in \Delta : \tau \cap \sigma = \varnothing\} \cup (\bar{v} * \partial\bar{\sigma} * \mathrm{lk}(\sigma, \Delta)),$$

where $\bar{v}$ is a new vertex. If $e = \{u, v\} \in \Delta$, then we define the *edge contraction of Δ along e* to be

$$\mathrm{contr}(e, \Delta) = \{\tau \in \Delta : u \notin \tau\} \cup \{(\tau \cup v) \setminus u : u \subset \tau \in \Delta\}.$$

A $(d-1)$-dimensional simplicial complex Δ is called a *simplicial $(d-1)$-manifold* (or *triangulated $(d-1)$-manifold*) if its geometric realization $\|\Delta\|$ is homeomorphic

to a manifold. A *combinatorial* $(d-1)$*-manifold* (or *PL* $(d-1)$*-manifold*) is a simplicial complex such that every vertex link is PL homeomorphic to the boundary of a $(d-1)$-simplex or the $(d-2)$-simplex. In particular, if Δ is a combinatorial manifold, then the *boundary complex* of Δ, denoted as $\partial\Delta$, consists of the empty face and the faces whose links are combinatorial balls. The boundary complex of a combinatorial d-ball is a combinatorial $(d-1)$-sphere. The faces that are not in the boundary complex are called the *interior faces*. In dimension $d-1=2$ or 3, the class of combinatorial manifolds and the class of simplicial manifolds are the same. However, there exist non-PL triangulations of the $(d-1)$-sphere on $d+12$ vertices for all $d-1 \geq 5$; see [Björner and Lutz 2000, Theorem 7].

One advantage of working in the class of combinatorial manifolds can be explained by the following elegant theorem [Alexander 1930].

Theorem 2.1 (Alexander). *Two closed combinatorial manifolds are piecewise linear homeomorphic if and only if there exists a finite sequence of edge subdivisions and their inverses leading from one combinatorial manifold to the other.*

It is well known that if an edge $e=\{a,b\}$ in a closed combinatorial manifold Δ satisfies the *link condition* $\mathrm{lk}(e,\Delta)=\mathrm{lk}(a,\Delta)\cap\mathrm{lk}(b,\Delta)$, then the edge contraction $\mathrm{contr}(e,\Delta)$ preserves the PL type of Δ; see [Nevo 2007].

There are two $(d-1)$-dimensional sphere bundles over the circle. We denote by $\mathbb{S}^{d-2}\times\mathbb{S}^1$ the orientable sphere bundle and by $\mathbb{S}^{d-2}\,\tilde{\times}\,\mathbb{S}^1$ the nonorientable sphere bundle. If M is a $(d-1)$-dimensional manifold, then we write $\#_i M$ as the (topological) connected sum of i copies of M. The connected sum of simplicial complexes is defined as an analog of the topological connected sum: let Δ and Γ be pure $(d-1)$-dimensional simplicial complexes and let $\sigma\in\Delta$, $\tau\in\Gamma$ be facets. The connected sum of Δ and Γ, $\Delta\#_{\sigma\sim\tau}\Gamma$ (or simply $\Delta\#\Gamma$, if σ and τ are understood in the context), is the complex obtained from $\Delta\cup\Gamma$ by removing σ,τ and then identifying $\partial\bar{\sigma}$ with $\partial\bar{\tau}$.

Let $\chi(\Delta)=\sum_{i=0}^{d-1}(-1)^i\beta_i(\Delta)$ be the *Euler characteristic* of the $(d-1)$-dimensional complex Δ, where $\beta_i(\Delta)$ is the rank of the i-th homology group of Δ computed with coefficients in $\mathbb{Z}$. By the Euler characteristic formula, $\chi(\Delta)=\sum_{i=0}^{d-1}(-1)^i f_i(\Delta)$. Hence the f-vector of a triangulated surface Δ can be expressed as

$$f(\Delta)=\big(1,\, f_0(\Delta),\, 3(f_0(\Delta)-\chi(\Delta)),\, 2(f_0(\Delta)-\chi(\Delta))\big).$$

The following classification theorem of surfaces will come in handy in our discussion of flag triangulated surfaces; see [Seifert and Threlfall 1980].

Theorem 2.2 (classification theorem of surfaces). *Every closed and connected surface is homeomorphic to either a sphere, a connected sum of tori, or a connected sum of projective planes.*

A graph is an ordered pair $G = (V, E)$, such that V is a finite set, and $E \subseteq \binom{V}{2}$. The graph of a simplicial complex Δ is $G(\Delta) = (V(\Delta), E(\Delta))$. A simplicial complex Δ is *flag* if all minimal nonfaces of Δ, also called *missing faces*, have cardinality 2; equivalently, Δ is the clique complex of $G(\Delta)$. For example, let $C_d^* = \mathrm{conv}\{\pm e_1, \ldots, \pm e_d\}$ be the d-cross-polytope, where the e_i's form the standard basis of $\mathbb{R}^d$. The boundary complex ∂C_d^* is a flag $(d-1)$-sphere. The properties of flag complexes are described in the following lemma.

Lemma 2.3 [Nevo and Petersen 2011, Lemma 5.2]. *Let Δ be a flag complex on vertex set V:*

(1) *If $W \subseteq V(\Delta)$, then $\Delta[W]$ is also flag.*

(2) *If σ is a face in Δ, then $\mathrm{lk}(\sigma) = \Delta[V(\mathrm{lk}(\sigma))]$. In particular, all links in a flag complex are also flag.*

(3) *If $W \subset V(\Delta)$, then $\|\Delta\| - \|\Delta[W]\|$ deformation retracts onto $\|\Delta[V - W]\|$.*

(4) *Any edge $\{v, v'\}$ in Δ satisfies the link condition $\mathrm{lk}(v) \cap \mathrm{lk}(v') = \mathrm{lk}(\{v, v'\})$.*

The flag analog of Alexander's theorem is the following theorem; see [Lutz and Nevo 2016].

Theorem 2.4 (Lutz–Nevo). *Two flag simplicial complexes are piecewise linearly homeomorphic if and only if they can be connected by a sequence of flag complexes, each obtained from the previous one by either an edge subdivision or edge contraction.*

3. Algorithms and results

It is well known that the minimal flag triangulation of $\mathbb{S}^2$ is the octahedral 2-sphere with six vertices. Our goal in this section is to find the minimal flag triangulations of three other surfaces: the torus, the Klein bottle and the real projective plane. In the next section we will discuss how to generate small flag triangulations of all types of surfaces based on triangulations of these three surfaces.

Our implementation is based on the Lutz–Nevo theorem. Since in dimension $d = 2, 3$, the class of combinatorial d-manifolds is the same as the class of triangulated d-manifolds, Theorem 2.4 guarantees that the minimal flag triangulation of a surface M can be obtained by applying edge subdivisions and admissible edge contractions on a given (possibly very large) flag triangulation of M. In an edge subdivision of a surface, a chosen edge e of the simplicial complex is divided into two edges, with the facets containing e replaced by four facets containing the two new edges. The resulting complex is always flag. In an edge contraction, we choose an edge $e = \{a, b\}$ and identify b with a; in other words, the edge e is contracted to the vertex a. Faces of the form $F \cup \{b\}$ are replaced by the faces $F \cup \{a\}$. However,

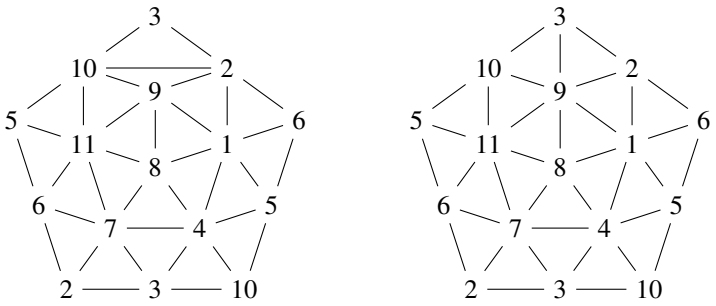


Figure 1. Two 11-vertex flag triangulations of $\mathbb{R}P^2$.

5 6 11	6 7 11	7 8 11	8 9 11	9 10 11
5 10 11	1 2 9	2 9 10	2 3 10	3 4 10
4 5 10	1 4 5	1 5 6	1 2 6	2 6 7
2 3 7	3 4 7	4 7 8	1 4 8	1 8 9

5 6 11	6 7 11	7 8 11	8 9 11	9 10 11
5 10 11	1 2 9	2 3 9	3 9 10	3 4 10
4 5 10	1 4 5	1 5 6	1 2 6	2 6 7
2 3 7	3 4 7	4 7 8	1 4 8	1 8 9

Table 1. Facets of two flag 11-vertex triangulations of $\mathbb{R}P^2$.

edge contraction does not always preserve flagness. The resulting complex is flag if and only if the edge contracted is not contained in any induced 4-cycle of the original simplicial complex; see [Lutz and Nevo 2016, Corollary 6.2]. We call such an edge an *admissible* edge.

Our computer search algorithm is as follows: we first build a relatively small flag triangulation of a given surface and apply a random sequence of edge subdivisions on the complex until the number of vertices reaches a set number. Then a sequence of admissible edge contractions is performed until a local minimum on f_0 is attained; i.e., no more admissible edges exist in the complex. The above process of edge subdivisions followed by edge contractions is iterated a given number of times. In each iteration, we keep track of the number of vertices in the complexes.

In what follows, we summarize the smallest flag triangulations of $\mathbb{R}P^2$, $\mathbb{S}^1 \times \mathbb{S}^1$ and $\mathbb{S}^1 \tilde{\times} \mathbb{S}^1$ found by our implementation.

Example 3.1 (triangulation of $\mathbb{R}P^2$). We started with an 11-vertex flag triangulation of $\mathbb{R}P^2$ as shown in the left of Figure 1. Our program found another nonisomorphic construction of $\mathbb{R}P^2$ with 11 vertices (see the right of Figure 1). Observe that, as the picture suggests, the right-hand triangulation has automorphism group Dih_5 (the symmetry group of the pentagon), while the left-hand triangulation has a different automorphism group. The facets of these two triangulations are listed in Table 1. In particular, they differ by one bistellar flip, i.e., by replacing the 2-faces $\{2, 3, 10\}$, $\{2, 9, 10\}$ with $\{2, 3, 9\}$, $\{3, 9, 10\}$. We will prove in the next section that the minimal flag triangulations of $\mathbb{R}P^2$ indeed have 11 vertices.

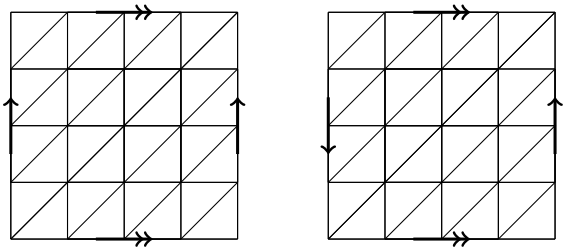


Figure 2. The 16-vertex flag triangulations of the torus (left) and the Klein bottle (right).

1 2 3	1 2 5	1 3 11	1 4 5	1 4 12	1 11 12
2 3 8	2 5 6	2 6 7	2 7 8	3 8 9	3 9 10
3 10 11	4 5 10	4 7 8	4 7 10	4 8 12	5 6 9
5 9 10	6 7 11	6 9 12	6 11 12	7 10 11	8 9 12

Table 2. Facets of the flag 12-vertex triangulation of $\mathbb{S}^1 \times \mathbb{S}^1$.

Example 3.2 (triangulation of the torus). We started with a flag triangulation of the torus with 16 vertices as shown in the left of Figure 2. A unique flag triangulation with 12 vertices was found in the searching process. The facets are listed in Table 2. We will show in the next section that this is indeed the unique minimal flag triangulation of the torus; see Figure 5.

Example 3.3 (triangulation of the Klein bottle). We started the searching process with a flag triangulation of the Klein bottle with 16 vertices as shown in Figure 2. Our program suggested that a minimal flag triangulation of the Klein bottle has 14 vertices, and there are at least 28 nonisomorphic such triangulations. We will see in the next section how to obtain several 14-vertex triangulations from 11-vertex flag triangulations of $\mathbb{R}P^2$.

Remark 3.4. The minimal flag triangulation of $\mathbb{S}^1 \times \mathbb{S}^1$ is exactly the vertex-transitive 2-manifold ${}^2 12_1^{83}$ found in [Lutz 2011]. It admits the group action of $S_4 \times S_3$.

4. Proof of minimality

In Section 3, we saw that there exist flag triangulations of $\mathbb{R}P^2$ and $\mathbb{S}^1 \times \mathbb{S}^1$ with 11 and 12 vertices, respectively. In this section, we prove that indeed our constructions give the minimal flag triangulations. The following lemma provides a necessary condition for a flag triangulation to be vertex-minimal.

Lemma 4.1. *Let Δ be a minimal flag triangulated surface. Then every edge is in an induced 4-cycle of Δ.*

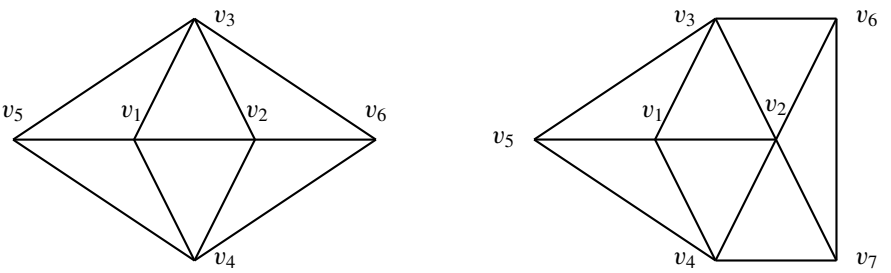


Figure 3. Links in the proofs of Lemma 4.2 (left), where $\deg v_1 = \deg v_2 = 4$, and Lemma 4.3 (right), where $\deg v_1 = 4$, $\deg v_2 = 5$.

Proof. Suppose not; then there is an edge $e = \{u, v\} \in \Delta$, such that it is not in any induced 4-cycle of Δ. Then $\{u, v\}$ can be admissibly contracted into Δ', where Δ' is homeomorphic to Δ and Δ' contains no induced 3-cycle; i.e., Δ is flag. This contradicts that Δ is the minimal flag triangulation. □

The minimal flag triangulation of $\mathbb{R}P^2$. In what follows, by $(w_1, w_2, \ldots, w_n)$ we denote the n-cycle with edges $\{w_i, w_{i+1}\}$ for $1 \le i \le n-1$ and $\{w_n, w_1\}$.

Lemma 4.2. *Let Δ be a minimal flag triangulation of a surface M. If there is an edge $e = \{v_1, v_2\} \in \Delta$ such that both of the links $\mathrm{lk}(v_1)$, $\mathrm{lk}(v_2)$ are 4-cycles, then $M \cong \mathbb{S}^2$ and Δ is the octahedral sphere.*

Proof. Assume that $\mathrm{lk}(v_1) = (v_5, v_3, v_2, v_4)$ and $\mathrm{lk}(v_2) = (v_1, v_3, v_6, v_4)$, as shown in the left of Figure 3. By Lemma 2.3, $\mathrm{lk}(v_1) \cap \mathrm{lk}(v_2) = \mathrm{lk}(\{v_1, v_2\})$ is the union of two vertices v_3, v_4 and hence v_5, v_6 are distinct. Since Δ is a minimal flag triangulation, the edge $\{v_5, v_1\}$ must be in at least one induced 4-cycle C of Δ. Since C is induced, $\{v_3, v_1\}$ or $\{v_4, v_1\}$ cannot be in C. So the edge $\{v_1, v_2\}$ is in C. Similarly, we have $\{v_2, v_6\} \subset C$. Hence $C = (v_5, v_1, v_2, v_6)$. Note that $G(\mathrm{st}(v_1) \cup \mathrm{st}(v_2) \cup \{v_5, v_6\})$ is the graph of an octahedral sphere. By flagness, this octahedral sphere is a subcomplex of Δ, and hence Δ must be the octahedral sphere. □

We can strengthen Lemma 4.2 as follows:

Lemma 4.3. *Let Δ be a minimal flag triangulation of a surface M. Then no adjacent vertices can have degree 4 and 5, respectively, in Δ.*

Proof. Assume that $\mathrm{lk}(v_1) = (v_5, v_3, v_2, v_4)$ and $\mathrm{lk}(v_2) = (v_1, v_3, v_6, v_7, v_4)$, as shown in the right of Figure 3. Using the same proof as in Lemma 4.2, v_5, v_6, v_7 are distinct vertices and the induced 4-cycle C that contains the edge $\{v_5, v_1\}$ must also contain $\{v_1, v_2\}$. By symmetry and without loss of generality, we may assume that $\{v_2, v_6\}$ is in C. Hence $C = (v_5, v_1, v_2, v_6)$. By flagness $\mathrm{lk}(v_3)$ must be the 4-cycle C. Now both $\mathrm{lk}(v_3)$ and $\mathrm{lk}(v_1)$ are 4-cycles and $\{v_1, v_3\} \in \Delta$. By Lemma 4.2, Δ is the octahedral sphere. This contradicts that $\mathrm{lk}(v_2)$ is a 5-cycle. □

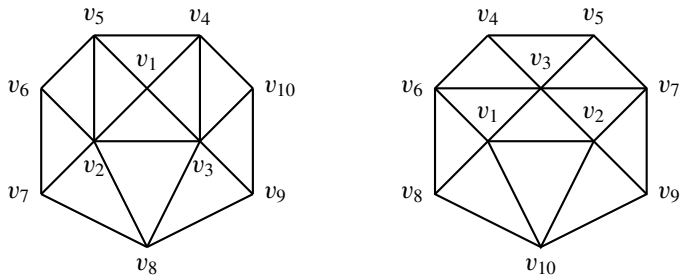


Figure 4. Links for case 1 of the proof of Theorem 4.4 (left), where $(\deg v_1, \deg v_2, \deg v_3) = (4, 6, 6)$, and case 2 (right), where $(\deg v_1, \deg v_2, \deg v_3) = (5, 5, 6)$.

Now we are ready to prove our main theorem.

Theorem 4.4. *Let Δ be a minimal flag triangulation of a surface $M \neq \mathbb{S}^2$. Then $f_0(\Delta) \geq 11$.*

Proof. Let $F = \{v_1, v_2, v_3\} \in \Delta$. By Lemmas 4.2 and 4.3, $(\deg v_1, \deg v_2, \deg v_3)$ is equal to either $(4, \geq 6, \geq 6)$ or $(\geq 5, \geq 5, \geq 5)$.

Case 1: The vertex v_1 is of degree 4. Since $\mathrm{lk}(v_i) \cap \mathrm{lk}(v_j) = \mathrm{lk}(\{v_i, v_j\})$ consists of two distinct vertices for any distinct $i, j \in \{1, 2, 3\}$, by the inclusion-exclusion principle and flagness of Δ we have

$$f_0(\Delta) \geq f_0\left(\bigcup_{i=1}^{3} \mathrm{lk}(v_i)\right) = \sum_{i=1}^{3} f_0(\mathrm{lk}(v_i)) - \sum_{1 \leq i < j \leq 3} f_0(\mathrm{lk}(v_i) \cap \mathrm{lk}(v_j))$$
$$= \sum_{i=1}^{3} f_0(\mathrm{lk}(v_i)) - 6 \geq 10.$$

In particular, if $f_0(\Delta) = 10$, then it follows that $(\deg v_1, \deg v_2, \deg v_3) = (4, 6, 6)$. Assume that the links of v_1, v_2, v_3 are as shown in the left of Figure 4. Since $\deg v_1 = 4$, by Lemma 4.3 $\mathrm{lk}(v_4)$ is not a 4-cycle and hence $\{v_5, v_{10}\} \notin \Delta$. Since $\mathrm{lk}(v_2)$ is flag, $\{v_5, v_3\}$, $\{v_5, v_7\}$ and $\{v_5, v_8\}$ are not edges of Δ. Hence v_5 can have at most degree 5 in Δ. However by Lemma 4.2, $\deg v_5 \geq 6$, a contradiction. Therefore, $f_0(\Delta) \geq 11$.

Case 2: Every vertex in Δ is of degree at least 5. It follows from the Euler characteristic formula that $f(\Delta) = (1, f_0(\Delta), 3(f_0(\Delta) - \chi(\Delta)), 2(f_0(\Delta) - \chi(\Delta)))$; here $\chi(\Delta) \leq 1$ as $M \neq \mathbb{S}^2$. If all vertices of Δ are of degree 5, by double counting we have

$$6(f_0(\Delta) - \chi(\Delta)) = 2f_1(\Delta) = \sum_{v \in \Delta} f_0(\mathrm{lk}(v)) = 5f_0(\Delta).$$

That is, $f_0(\Delta) = 6\chi(\Delta) \leq 6$, a contradiction.

Without loss of generality, assume $(\deg v_1, \deg v_2, \deg v_3) = (\geq 5, \geq 5, \geq 6)$. As before, we have $f_0(\Delta) \geq f_0\big(\bigcup_{i=1}^3 \mathrm{lk}(v_i)\big) \geq 5+5+6-6 = 10$. Suppose $f_0(\Delta) = 10$, in which case $(\deg v_1, \deg v_2, \deg v_3) = (5, 5, 6)$ and the links of v_i are as shown in the right of Figure 4. By Lemma 4.3, $\mathrm{lk}(v_{10})$ and $\mathrm{lk}(v_6)$ are not 4-cycles and hence $\{v_8, v_9\}, \{v_8, v_4\} \notin \Delta$. Since $\deg(v_8) \geq 5$, we must have $\{v_8, v_5\}, \{v_8, v_7\} \in \Delta$ and $\mathrm{lk}(v_8)$ is equal to $(v_6, v_1, v_{10}, v_5, v_7)$ or $(v_6, v_1, v_{10}, v_7, v_5)$. However, both $\mathrm{lk}(v_2)$ and $\mathrm{lk}(v_3)$ are flag and $\{v_7, v_{10}\}, \{v_6, v_7\}$ cannot be edges of Δ. This leads to a contradiction. □

In Section 3 we showed the existence of an 11-vertex flag triangulation of $\mathbb{R}P^2$. We immediately obtain the following:

Theorem 4.5. *A minimal flag triangulation of $\mathbb{R}P^2$ has 11 vertices.*

The minimal flag triangulation of $\mathbb{S}^1 \times \mathbb{S}^1$. In what follows, we show that the minimal flag triangulation of $\mathbb{S}^1 \times \mathbb{S}^1$ has 12 vertices and furthermore, it is unique.

Lemma 4.6. *Let Δ be a minimal flag triangulation of $\mathbb{S}^1 \times \mathbb{S}^1$. Then $f_0(\Delta) \leq 12$ and every vertex of Δ is of degree ≤ 6.*

Proof. We know that $f_0(\Delta) \leq 12$, since in Section 3 we found a flag triangulation of $\mathbb{S}^1 \times \mathbb{S}^1$ with 12 vertices. Assume that there is a vertex v of degree ≥ 7. Let $W = V(\mathrm{st}(v))$ and $W^c = V(\Delta) - W$. By Lemma 2.3, $\|\Delta - \Delta[W]\| = \|\Delta - \mathrm{st}(v)\|$ deformation retracts onto $\|\Delta[W^c]\|$. On the other hand, since $|W^c| = |V(\Delta)| - |W| \leq 12 - 8 \leq 4$ and $\Delta[W^c]$ is flag, it follows that either $\Delta[W^c]$ is the 4-cycle or it is contractible. Hence

$$2 = \beta_1(\|\Delta - v\|) = \beta_1(\|\Delta - \mathrm{st}(v)\|) = \beta_1(\|\Delta[W^c]\|) \leq 1,$$

a contradiction. □

Lemma 4.7. *A minimal flag triangulation Δ of $\mathbb{S}^1 \times \mathbb{S}^1$ has 12 vertices and each vertex is of degree 6.*

Proof. Since $\chi(\Delta) = 0$, by the Euler characteristic, $f(\Delta) = (1, f_0, 3f_0, 2f_0)$. By Lemma 4.6,

$$6f_0(\Delta) = 2f_1(\Delta) = \sum_{v \in \Delta} f_0(\mathrm{lk}(v)) \leq 6f_0(\Delta).$$

Hence $f_0(\mathrm{lk}(v)) = 6$ for every vertex of Δ. If $\{v_1, v_2, v_3\}$ is a 2-face of Δ, then

$$\begin{aligned} 12 \geq f_0(\Delta) &\geq f_0\left(\bigcup_{i=1}^3 \mathrm{lk}(v_i)\right) \\ &\geq \sum_{i=1}^3 f_0(\mathrm{lk}(v_i)) - \sum_{1 \leq i < j \leq 3} f_0(\mathrm{lk}(\{v_i, v_j\}) = 3 \cdot 6 - 3 \cdot 2 = 12. \end{aligned}$$

This proves the claim. □

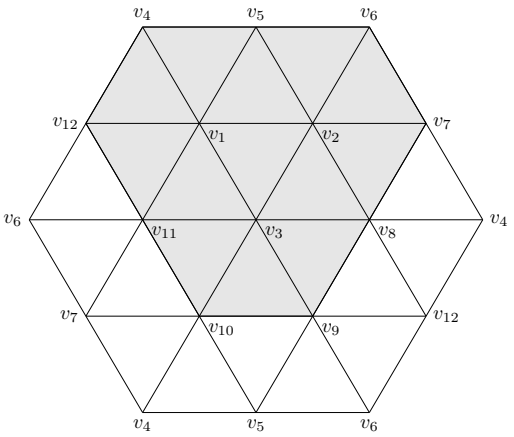


Figure 5. The triangulation of $\mathbb{S}^1 \times \mathbb{S}^1$ obtained by identifying three pairs of (oriented) opposite sides in the triangulated hexagon.

Theorem 4.8. *The minimal flag triangulation of* $\mathbb{S}^1 \times \mathbb{S}^1$ *has* 12 *vertices and is unique.*

Proof. Let $\{v_1, v_2, v_3\}$ be a 2-face of Δ. By the above lemma, $\left|V\left(\bigcup_{i=1}^3 \mathrm{st}(v_i)\right)\right| = |V(\Delta)| = 12$. Furthermore, $G(\Delta)$ contains the subgraph $G\left(\bigcup_{i=1}^3 \mathrm{st}(v_i)\right)$, as shown in the shaded part of Figure 5.

Since $\mathrm{lk}(v_i) = \Delta[V(\mathrm{lk}(v_i))]$, by flagness v_5 is not connected to $v_7, v_8, v_3, v_{11}, v_{12}$. On the other hand, $\deg(v_5) = 6$. Hence v_5 must be connected to v_9, v_{10}. Applying the same argument to v_8, v_{11}, we have $\{v_8, v_4\}, \{v_8, v_{12}\}, \{v_{11}, v_6\}, \{v_{11}, v_7\} \in \Delta$.

It is left to decide the remaining six edges. Since v_4 is of degree 6, v_4 is connected to two vertices among v_7, v_9, v_{10}. However, if both $\{v_4, v_9\}$ and $\{v_4, v_{10}\}$ are edges of Δ, $\{v_4, v_5, v_9, v_{10}\}$ will form a clique in $G(\Delta)$ and by flagness of Δ form a 3-face of Δ, which is not possible. Hence $\{v_4, v_7\} \in \Delta$. Similarly, we apply the same argument on the vertices $v_7, v_{10}, v_6, v_9, v_{12}$ respectively. This shows that $\{v_7, v_{10}\}$, $\{v_{10}, v_4\}$, $\{v_6, v_9\}$, $\{v_9, v_{12}\}$ and $\{v_{12}, v_6\}$ are edges of Δ. This gives all 36 edges in Δ and the graph is isomorphic to the 12-vertex construction in Section 3. □

Generating small flag triangulations of all surfaces. By the classification theorem of surfaces (see Theorem 2.2), a surface is either a 2-sphere, or the connected sum of $\mathbb{S}^1 \times \mathbb{S}^1$, or the connected sum of $\mathbb{R}P^2$. Hence to construct a small flag triangulation of all surfaces, it suffices to show how to take the connected sum of surfaces efficiently while preserving flagness.

Definition 4.9. Let Δ_1 and Δ_2 be the flag triangulations of $(d-1)$-manifolds M_1 and M_2, respectively. Let $W_1 \subset V(\Delta_1)$ and $W_2 \subset V(\Delta_2)$. Assume that $\Delta_1[W_1]$ and $\Delta_2[W_2]$ are simplicial $(d-1)$-balls and that $\partial\Delta_1[W_1]$ and $\partial\Delta_2[W_2]$ are flag

simplicial $(d-2)$-spheres; furthermore, assume that $\phi : \partial\Delta_1[W_1] \to \partial\Delta_2[W_2]$ is a simplicial isomorphism. Define a simplicial complex $\Delta = \Delta_1 \#_\phi \Delta_2$ by

(1) removing the interior faces of $\Delta_1[W_1]$ and $\Delta_2[W_2]$, and

(2) gluing Δ_1 and Δ_2 by identifying $\partial(\Delta_1[W_1])$ with $\phi(\partial(\Delta_1[W_1])) = \partial(\Delta_2[W_2])$.

We say Δ is a *flag connected sum* of Δ_1 and Δ_2 under ϕ.

Note that $\Delta_1[W_1]$ and $\Delta_2[W_2]$ are not necessarily isomorphic in Definition 4.9. The following lemma shows that Definition 4.9 generates a new flag complex.

Lemma 4.10. *If Δ_1 and Δ_2 are flag triangulations of $(d-1)$-manifolds M_1 and M_2, respectively, then the flag connected sum $\Delta = \Delta_1 \#_\phi \Delta_2$ is a flag triangulation of $M_1 \# M_2$.*

Proof. First Δ triangulates $M_1 \# M_2$ because $\|\Delta_1\| = M_1$, $\|\Delta_2\| = M_2$ and $\Delta_1[W_1]$ and $\Delta_2[W_2]$ are simplicial full-dimensional balls. Assume that Δ is not flag and F is a missing face of Δ of dimension ≥ 2. If there are two vertices $v_1, v_2 \in F$ such that $v_i \in \Delta_i$ but $v_i \notin W_i$ for $i = 1, 2$, then by the construction, v_1 is not connected to v_2 in Δ, contradicting the fact that F is a missing face of dimension ≥ 2. So all the vertices of F are either in Δ_1 or Δ_2; without loss of generality assume that they are in Δ_1. By flagness of Δ_1, we have $F \in \Delta_1$. But since $F \notin \Delta$, F must be an interior face of $\Delta_1[W_1]$ which is removed in the flag connected sum. In particular, F is a missing face of $\partial\Delta_1[W_1]$, contradicting the flagness of $\partial\Delta_1[W_1]$. □

The above lemma together with the minimal flag triangulations of $\mathbb{S}^1 \times \mathbb{S}^1$ and $\mathbb{R}P^2$ lead to the following upper bound on the number of vertices of minimal flag triangulated surfaces.

Theorem 4.11. *Let $k \geq 1$ be an integer. A minimal flag triangulation of the connected sum of k tori requires at most $8 + 4k$ vertices, while a minimal flag triangulation of the connected sum of k real projective planes requires at most $8 + 3k$ vertices.*

Proof. The proof is by induction on k. In the case of connected sum of tori, the base case is verified by the 12-vertex flag triangulation (whose vertex links are all 6-cycles) that we found in Section 3. Inductively, assume that there is a flag triangulation Δ_1 of the connected sum of $k-1$ tori with $4 + 4k$ vertices and, furthermore, that there is a vertex v_1 of degree 6 in Δ_1. Let Δ_2 be the 12-vertex flag triangulation of $\mathbb{S}^1 \times \mathbb{S}^1$. By the construction, there is a vertex v_2 of degree 6 in Δ_2. We let $W_i = V(\mathrm{st}(v_i, \Delta_i))$ for $i = 1, 2$. Then $\Delta_1[W_1] = \mathrm{st}(v_1, \Delta_1)$ and $\Delta_2[W_2] = \mathrm{st}(v_2, \Delta_2)$ are indeed simplicial 2-balls whose boundaries are both 6-cycles. Since the vertex stars and vertex links are induced, by Lemma 4.10 the flag connected sum $\Delta = \Delta_1 \# \Delta_2$ is well-defined and triangulates the connected

sum of k tori. Furthermore,

$$f_0(\Delta) = f_0(\Delta_1) + f_0(\Delta_2) - f_0(\mathrm{st}(v_1, \Delta_1)) - 1 = (4+4k) + 12 - 7 - 1 = 8 + 4k.$$

Finally, to make the inductive construction work, we need to check that there is a vertex w of degree 6 in the new complex Δ. Indeed, we can take any vertex w from Δ_2 such that $w \notin \mathrm{st}(v_2, \Delta_2)$. Any such w has $\mathrm{st}(w, \Delta_2) = \mathrm{st}(w, \Delta)$ and so it is of degree 6 in Δ.

The proof for connected sum of $\mathbb{R}P^2$ is similar. We begin with an 11-vertex flag triangulation of $\mathbb{R}P^2$ as in Figure 1 and note that it has two disjoint vertices of degree 6 (for example, the vertices 1 and 11). Inductively we take the flag connected sum of a $(5+3k)$-vertex flag triangulation Γ_1 of the connected sum of $k-1$ copies of $\mathbb{R}P^2$ found by induction with an 11-vertex flag triangulation Γ_2 of $\mathbb{R}P^2$. This is done by choosing a vertex u of degree 6 in Γ_1 (whose existence is by induction), removing all interior faces of $\mathrm{st}(u, \Gamma_1)$ and $\mathrm{st}(11, \Gamma_2)$, and identifying $\mathrm{lk}(u, \Gamma_1)$ with $\mathrm{lk}(11, \Gamma_2)$. The vertex 1 from Γ_2 is also a vertex of degree 6 in $\Gamma_1 \# \Gamma_2$. Finally,

$$\begin{aligned} f_0(\Gamma_1 \# \Gamma_2) &= f_0(\Gamma_1) + f_0(\Gamma_2) - f_0(\mathrm{st}(u, \Gamma_1)) - 1 \\ &= (5+3k) + 11 - 7 - 1 = 8 + 3k. \end{aligned}$$ □

Remark 4.12. As shown in Section 3, the computer search found 28 combinatorially distinct 14-vertex flag triangulations of the Klein bottle. Indeed many of these 14-vertex triangulations can be obtained by taking the flag connected sum of two 11-vertex flag triangulation of $\mathbb{R}P^2$, as suggested in the proof of Theorem 4.11.

5. Towards a manifold Charney–Davis conjecture

As mentioned in the Introduction, while the lower bound theorem for flag 3-spheres was established by Davis and Okun, so far there is no analogous lower bound conjecture for flag 3-manifolds. To motivate why such a conjecture might exist, let us recall the classical lower bound theorems for manifolds and balanced manifolds. A simplicial $(d-1)$-sphere is *stacked* if it is the connected sum of the boundaries of d-simplices. The *Walkup class* $\mathcal{H}^d$ $(d \geq 3)$ is defined recursively as follows:

(1) $\mathcal{H}^d(0)$ is the set of all stacked $(d-1)$-spheres, that is, the spheres obtained by successively taking the connected sum of the boundary complexes of the simplices.

(2) A simplicial complex Δ is in $\mathcal{H}^d(k+1)$ if it is obtained from a member of $\mathcal{H}^d(k)$ by a handle addition.

(3) $\mathcal{H}^d = \bigcup_{k \geq 0} \mathcal{H}^d(k)$.

Theorem 5.1 [Datta and Murai 2017; Murai 2015; Novik and Swartz 2009]. *For any connected simplicial $(d-1)$-manifold without boundary Δ and $d \geq 4$,*

$$g_2(\Delta) := f_1(\Delta) - df_0(\Delta) + \binom{d+1}{2} \geq \binom{d+1}{2}\beta_1(\Delta).$$

Furthermore, the equality is attained if and only if Δ is in the Walkup class.

We remark that a weaker inequality $g_2 \geq 0$ was first proved in the class of simplicial polytopes [Barnette 1973b] and in the class of simplicial manifolds [Barnette 1973a]. Using rigidity theory of frameworks, Kalai [1987] found an alternative proof of the inequality and also characterized all simplicial $(d-1)$-manifolds with $g_2 = 0$ assuming $d \geq 4$. In recent years, lower bound results were also established for balanced simplicial polytopes and manifolds. A $(d-1)$-dimensional simplicial complex Δ is called *balanced* if there is a coloring map $\kappa : V(\Delta) \to [d]$ such that if $\{a, b\} \in \Delta$, then $\kappa(a) \neq \kappa(b)$. The balanced Walkup class, $\mathcal{BH}^d$, consists of balanced $(d-1)$-dimensional complexes that are obtained from the boundary complexes of d-cross-polytopes by successively applying the operations of balanced connected sums and balanced handle additions.

Theorem 5.2 [Klee and Novik 2016; Juhnke-Kubitzke et al. 2018]. *For any connected balanced simplicial $(d-1)$-manifold without boundary Δ and $d \geq 4$,*

$$\bar{g}_2(\Delta) := 2f_1(\Delta) - 3(d-1)f_0(\Delta) + 2d(d-1) \geq 4\binom{d}{2}\beta_1(\Delta).$$

Furthermore, for $d \geq 5$ the equality holds if and only if Δ is in the balanced Walkup class.

In particular, the above theorems imply that if Δ is a simplicial $(d-1)$-manifold (balanced simplicial $(d-1)$-manifold, resp.) that attains the lower bound on g_2 ($\bar{g}_2$, resp.), then its geometric realization $\|\Delta\|$ could be (the connected sum of) sphere bundles over the circle but could not be a $(d-1)$-dimensional torus, $\mathbb{R}P^{d-1}$, etc.

As suggested in [Gal 2005], the analog of g-numbers for any flag $(d-1)$-manifold Δ are the γ-numbers. In what follows, we will only consider the first two γ-numbers of Δ, given by

$$\gamma_1(\Delta) = f_0(\Delta) - 2d, \quad \gamma_2(\Delta) = f_1(\Delta) - (2d-3)f_0(\Delta) + 2d(d-2).$$

The following statements are the celebrated Davis–Okun theorem [2001] and a special case of the γ-conjecture [Gal 2005] (which we will call the γ_2-conjecture). The γ_2-conjecture has been verified for several classes of flag simplicial spheres; see, for example, [Karu 2006; Nevo and Petersen 2011].

Theorem 5.3 [Davis and Okun 2001]. *Let Δ be a flag simplicial 3-sphere. Then $\gamma_2(\Delta) \geq 0$.*

Conjecture 5.4 [Gal 2005]. *Let Δ be a flag simplicial $(d-1)$-sphere. Then $\gamma_2(\Delta) \geq 0$.*

As we see from Theorems 5.1 and 5.2, the lower bound theorem for (balanced) simplicial spheres can be extended to (balanced) simplicial manifolds. A natural question to ask is: does there also exist a generalization of the Davis–Okun theorem to the class of flag simplicial 3-manifolds? Is there a reasonable manifold γ_2-conjecture?

Generating flag triangulations of 3-manifolds. In this section, we collect data on certain flag 3-manifolds Δ to test whether there is relation between the minimum $\gamma_2(\Delta)$ and $\beta_1(\Delta)$. The candidates of types of manifolds that could attain small γ_2 with respect to β_1 are (the connected sum) of sphere bundles over the circle. To obtain a flag triangulation of a 3-manifold M, one can always take the barycentric subdivision of any triangulation of M. However, taking barycentric subdivisions usually generates a lot of new vertices. In what follows we introduce another useful way to construct flag triangulations of products of manifolds, and discuss how to apply the flag connected sum. This applies to finding small flag triangulations of $\#_i\, \mathbb{S}^2 \times \mathbb{S}^1$, $\#_i\, \mathbb{S}^2 \,\widetilde{\times}\, \mathbb{S}^1$ and $\#_i\, \mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{S}^1$.

Step 1: Construct the flag triangulations of products. The method of constructing triangulations of products is well known; see, for example, [Lee 1997; Eilenberg and Steenrod 1952]. We follow the description in [Lutz 2003, Section 3]. Let Δ_1 and Δ_2 be the flag triangulations of two manifolds M_1 and M_2, respectively. A cell decomposition of $M_1 \times M_2$ is given by taking the union of all cells $\sigma_m \times \sigma_n$, where $\sigma_m = \{u_0, u_1, \ldots, u_m\}$ is an m-face of Δ_1 and $\sigma_n = \{v_0, v_1, \ldots, v_n\}$ is an n-face of Δ_2. Identify the vertices $\{(u_i, v_j) : 0 \le i \le m,\ 0 \le j \le n\}$ with the points $\{(i, j) : 0 \le i \le m,\ 0 \le i \le n\}$ in $\mathbb{Z}^2$. Then the set of all monotone increasing lattice paths $(u_{i_0} = u_0, v_{i_0} = v_0) - (u_{i_1}, v_{i_1}) - \cdots - (u_{i_{m+n}} = u_m, v_{i_{m+n}} = v_n)$ corresponds to the set of facets $\{(u_{i_0}, v_{i_0}), \ldots, (u_{i_{m+n}}, v_{i_{m+n}})\}$ of $\sigma_m \times \sigma_n$. This is called the *staircase triangulation of product of simplices*. Furthermore, if the vertices of Δ_1 and Δ_2 are totally ordered, then the union of all faces of the staircase triangulation of $\sigma_m \times \sigma_n$ gives a *consistent* product triangulation of $M_1 \times M_2$. We call it the *staircase triangulation of* $\Delta_1 \times \Delta_2$.

Lemma 5.5. *If Δ_1 and Δ_2 are flag triangulations of the manifolds M_1 and M_2, respectively, then the staircase triangulation Δ of $\Delta_1 \times \Delta_2$ is also flag.*

Proof. Assume that the vertices of Δ_1 and Δ_2 are totally ordered as $(u_0, u_1, \ldots)$ and $(v_0, v_1, \ldots)$, respectively. Suppose F is a missing face of Δ of size $k > 2$.

First we claim that there is an order of $V(F) = \{s_1, s_2, \ldots, s_k\}$ where $s_i = (u_{p_i}, v_{q_i})$ such that $p_1 \le p_2 \le \cdots \le p_k$ and $q_1 \le q_2 \le \cdots \le q_k$. Indeed, if for some $i < j$ we have $p_i < p_j$ and $q_i > q_j$ (or similarly, $p_i > p_j$ and $q_i < q_j$), then the points (p_i, q_i) and (p_j, q_j) do not belong to any monotone increasing lattice path in $\mathbb{Z}^2$, and hence $\{s_i, s_j\} \notin \Delta$, a contradiction.

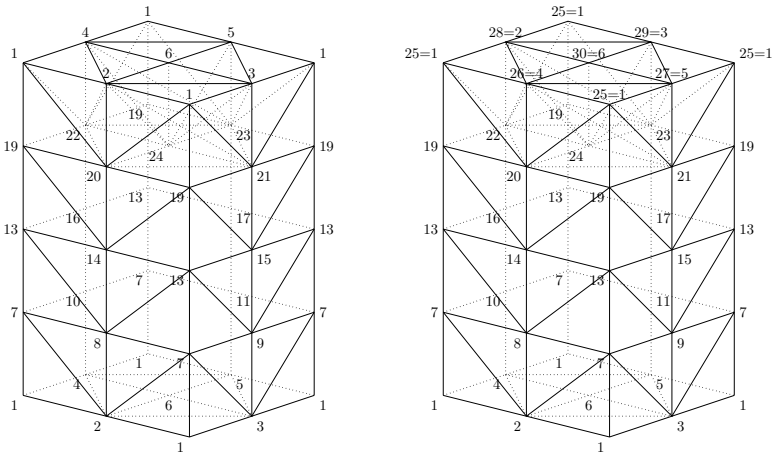


Figure 6. A triangulation of the orientable and nonorientable 3-dimensional sphere bundles over $\mathbb{S}^1$, respectively. Only the edges in the top layer of the prism are shown for simplicity.

By the definition of the staircase triangulation, any pair of distinct u_{p_i} forms an edge in Δ_1. Hence, $P = \{u_{p_1}, u_{p_2}, \ldots, u_{p_k}\}$ is a clique in Δ_1. Since Δ_1 is flag, we have $P \in \Delta_1$. Similarly, $Q = \{v_{q_1}, v_{q_2}, \ldots, v_{q_k}\} \in \Delta_2$. In other words, all the vertices of F are in the cell $P \times Q$. Since the vertices of F are in an increasing order in both u and v, they are also lattice points on some monotone increasing lattice path from (p_1, q_1) to (p_k, q_k). Hence F is a face in the staircase triangulation of $P \times Q$, i.e., $F \in \Delta$, which leads to a contradiction. □

As an illustration, to form flag triangulations of $\mathbb{S}^2 \times \mathbb{S}^1$ (orientable) and $\mathbb{S}^2 \,\tilde{\times}\, \mathbb{S}^1$ (nonorientable), we take the flag triangulation of $\mathbb{S}^2$ as the octahedral 2-sphere and the flag triangulation of $\mathbb{S}^1$ as the 4-cycle. Based on the above lemma, we generate two triangulations as shown in Figure 6.

Step 2: Apply the flag connected sum. To generate the flag connected sum of flag 3-manifolds, we identify isomorphic edge stars and then remove all the faces containing the corresponding edges. (We use the fact that the edge stars are induced subcomplexes in the flag complexes whose boundaries are also flag.) One advantage of applying this special flag connected sum is that it does little change to the graphs of the complexes.

Step 3: Search for minimal γ_2 by the Lutz–Nevo theorem. We generate a few flag triangulations of 3-manifolds by using either the staircase triangulation or the barycentric subdivision of the minimal (nonflag) triangulation. From the Lutz–Nevo theorem and our implementation as described in Section 3, we obtain data on estimated minimum γ_2 for various triangulated 3-manifolds, summarized in Table 3.

	β_1	minimum γ_2
$\mathbb{S}^3$	0	0
$\mathbb{R}P^3$	0	38
$L(3,1)$	0	82
$\#_i(\mathbb{S}^2 \times \mathbb{S}^1)$ for $i \in [10]$	i	$16i$
$\#_i(\mathbb{S}^2 \,\widetilde{\times}\, \mathbb{S}^1)$ for $i \in [10]$	i	$16i$
$\#_i(\mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{S}^1)$ for $i = 1, 2$	$3i$	$112i$
$\#_i\, \mathbb{S}^3 \times \mathbb{S}^1$ for $i \in [4]$	i	$30i$

Table 3. Estimated minimal γ_2 for several flag 3- and 4-manifolds; here $[n] = \{1, 2, \ldots, n\}$.

We propose the following 3-dimensional manifold Charney–Davis conjecture.

Conjecture 5.6. *Let Δ be a flag 3-manifold. Then $\gamma_2(\Delta) \geq 16\beta_1(\Delta)$.*

Several remarks are in order. First, it is not surprising that the estimated minimum γ_2 could be achieved by many distinct flag triangulations of a given manifold. Second, the data suggest that the minimum γ_2-numbers for flag triangulations of certain 3-manifolds are highly linear with respect to their first Betti numbers. Third, among the 3-manifolds having the same β_1, the connected sum of sphere bundles over the circle has smaller γ_2. Fourth, although the lower bound on γ_2 for higher dimensional flag manifolds is also of interest, our program is not efficient enough to get reliable estimation in these cases. For sake of completeness, we include the results on the connected sum of $\mathbb{S}^3 \times \mathbb{S}^1$ in Table 3. (It is possible that there is a more general conjecture $\gamma_2 \geq 2d(d-2)\beta_1$. For lack of evidence, we exclude it from our conjecture.)

We close by showing the lower bound on γ_2 in the conjecture, if true, is tight. First we define the handle addition on flag complexes.

Definition 5.7. Let Δ be a pure flag simplicial complex of dimension $d-1$. For any two vertices u and v, let $\mathrm{dist}(u, v)$ be the length of a shortest path from u to v in Δ. Assume that there are two disjoint subsets $W_1, W_2 \subset V(\Delta)$ such that

(1) both $\Delta[W_1]$ and $\Delta[W_2]$ are simplicial $(d-1)$-balls,

(2) $\partial\Delta[W_1]$ and $\partial\Delta[W_2]$ are isomorphic flag simplicial $(d-2)$-spheres, where $\phi : \partial\Delta[W_1] \to \partial\Delta[W_2]$ is a simplicial isomorphism, and

(3) $\mathrm{dist}(u, v) \geq 4$ for every $u \in W_1$ and $v \in W_2$.

The simplicial complex Δ^ϕ obtained from Δ by removing all interior faces of $\Delta[W_1]$ and $\Delta[W_2]$ and identifying each $v \in W_1$ with $\phi(v)$ is called a *flag handle addition* to Δ.

Lemma 5.8. *If Δ is flag, then a flag handle addition Δ^ϕ is flag.*

Proof. Let $V = V(\Delta)$ and $\bar{\phi} : \Delta \to \Delta^{\phi}$ be the handle addition map. Assume that F is a missing face of Δ^{ϕ} of size > 2. If $\bar{\phi}^{-1}(F) \in \Delta$ contains a pair of vertices $u \in W_1$ and $v \in W_2$, it also contains a path from u to v in Δ. However, $\operatorname{dist}(u, v) \geq 4$, and hence F cannot be the clique on $V(F)$. Now assume that $V(F) \subseteq V \setminus W_2$. Since $\Delta[V \setminus W_2]$ is flag, we have $F \in \Delta[V \setminus W_2]$. Furthermore, $\Delta^{\phi}[W_1] = \partial\Delta[W_1]$ is flag implies that $F \in \Delta^{\phi}$, which leads to a contradiction. □

Lemma 5.9. *Let Δ_1 and Δ_2 be two flag 3-manifolds. Further assume that for two edges $e_1 \in \Delta_1$ and $e_2 \in \Delta_2$, there exists a simplicial isomorphism $\phi : \partial \operatorname{st}(e_1, \Delta_1) \to \partial \operatorname{st}(e_2, \Delta_2)$. Then*

$$\gamma_2(\Delta_1 \#_{\phi} \Delta_2) = \gamma_2(\Delta_1) + \gamma_2(\Delta_2) + 2\gamma_1(\operatorname{lk}(e_1, \Delta_1)).$$

Similarly, if Δ is a flag 3-manifold containing two edges σ_1, σ_2 such that there is a simplicial isomorphism $\psi : \partial \operatorname{st}(\sigma_1, \Delta) \to \partial \operatorname{st}(\sigma_2, \Delta)$ that defines a flag handle addition on Δ, then

$$\gamma_2(\Delta^{\psi}) = \gamma_2(\Delta) + 2\gamma_1(\operatorname{lk}(\sigma_1, \Delta)) + 16.$$

Proof. By the definition of γ_2,

$$\gamma_2(\Delta_1 \#_{\phi} \Delta_2) = f_1(\Delta_1 \#_{\phi} \Delta_2) - 5f_0(\Delta_1 \#_{\phi} \Delta_2) + 16.$$

Also by the definition of the flag connected sum,

$$\begin{aligned} f_1(\Delta_1 \#_{\phi} \Delta_2) &= f_1(\Delta_1) + f_1(\Delta_2) - f_1(\operatorname{st}(e_1, \Delta_1)) - 1, \\ f_0(\Delta_1 \#_{\phi} \Delta_2) &= f_0(\Delta_1) + f_0(\Delta_2) - f_0(\operatorname{st}(e_1, \Delta_1)). \end{aligned}$$

Combining the above equations, we have

$$\gamma_2(\Delta_1 \#_{\phi} \Delta_2) = \gamma_2(\Delta_1) + \gamma_2(\Delta_2) + (-f_1(\operatorname{st}(e_1, \Delta_1)) + 5f_0(\operatorname{st}(e_1, \Delta_1) - 17).$$

The last term on the right-hand side of the above equation is

$$\begin{aligned} &-f_1(\operatorname{st}(e_1)) + 5f_0(\operatorname{st}(e_1)) - 17 \\ &\qquad = -(f_0(\operatorname{lk}(e_1)) + 1) + 2f_0(\operatorname{lk}(e_1)) + 5(f_0(\operatorname{lk}(e_1)) + 2) - 17 \\ &\qquad = 2f_0(\operatorname{lk}(e_1)) - 8 = 2\gamma_1(\operatorname{lk}(e_1)), \end{aligned}$$

which proves the first claim. The proof of the second claim is similar:

$$\begin{aligned} \gamma_2(\Delta^{\psi}) &= (f_1(\Delta) - f_1(\operatorname{st}(e_1)) - 1) - 5(f_0(\Delta) - f_0(\operatorname{st}(e_1))) + 16 \\ &= \gamma_2(\Delta) - (3f_0(\operatorname{lk}(e_1)) + 1) - 1 + 5(f_0(\operatorname{lk}(e_1)) + 2) \\ &= \gamma_2(\Delta) + 2\gamma_1(\operatorname{lk}(e_1)) + 16. \end{aligned}$$

□

Proposition 5.10. *For every positive integer b, there exists a flag 3-manifold Δ with $\beta_1(\Delta) = b$ and $\gamma_2(\Delta) = 16b$.*

Proof. The construction is done in three steps.

Step 1: Construct a 3-manifold Δ_4 with two edges e and e' such that their links are 4-cycles, and the vertex sets of their stars are disjoint. Let $\Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4$ be four disjoint copies of the octahedral 3-sphere, and let $\Delta_1 = \Gamma_1$. We will inductively define a complex Δ_i for $2 \le i \le 4$ by $\Delta_i = \Delta_{i-1} \#_{\phi_{i-1}} \Gamma_i$ as follows. Assign colors 1, 2, 3, 4 to the vertices of each Γ_i by giving antipodal vertices the same color. Let e and e_1 be distinct edges of color 1, 2 in Δ_1, and let e'_1 be an edge of color 1, 2 in Γ_2. There is a color-preserving isomorphism $\phi_1 : \partial\,\mathrm{st}(e_1, \Delta_1) \to \partial\,\mathrm{st}(e'_1, \Gamma_2)$, and we define $\Delta_2 = \Delta_1 \#_{\phi_1} \Gamma_2$, which may be viewed as the join of the 4-cycle $\mathrm{lk}(e'_1, \Gamma_2)$ colored by 3, 4 and a 6-cycle colored by 1, 2.

Next choose $e_2 = \{v_1, v_2\} \in \Delta_2$, where v_1 is a vertex of color 3 in $\mathrm{lk}(e'_1, \Gamma_2)$ and v_2 is a vertex of color 2 not in Δ_1. Given an edge e'_2 of color 2, 3 in Γ_3, use a color-preserving isomorphism $\phi_2 : \partial\,\mathrm{st}(e_2, \Delta_2) \to \partial\,\mathrm{st}(e'_2, \Gamma_3)$ to define $\Delta_3 = \Delta_2 \#_{\phi_2} \Gamma_3$. Here we have that $\mathrm{lk}(e'_2, \Delta_3)$ is a cycle of colors 1, 4.

Then choose the edge $e_3 = \{v_3, v_4\} \in \Delta_3$ such that the vertex $v_3 \in \mathrm{lk}(e'_2, \Gamma_3)$ is of color 4 and the vertex $v_4 \notin \Delta_2$ is of color 3. Given an edge e'_3 of color 3, 4 in Γ_4, let e' be its antipodal edge in Γ_4, and use a color-preserving isomorphism $\phi_3 : \partial\,\mathrm{st}(e_3, \Delta_3) \to \partial\,\mathrm{st}(e'_3, \Gamma_4)$ to define $\Delta_4 = \Delta_3 \#_{\phi_3} \Gamma_4$. Then e' is disjoint from Δ_1 and does not share a color with e; thus we verify that the links of e and e' in Δ_4 are indeed disjoint.

Step 2: Construct a flag 3-manifold Δ_{16} from Δ_4 and apply flag handle addition. Take four copies $\Delta_4^1, \Delta_4^2, \Delta_4^3, \Delta_4^4$ of the above Δ_4 to form a flag connected sum

$$\Delta_{16} = \Delta_4^1 \#_{\psi_1} \Delta_4^2 \#_{\psi_2} \Delta_4^3 \#_{\psi_3} \Delta_4^4,$$

where $\psi_i : \partial\,\mathrm{st}(e', \Delta_4^i) \to \partial\,\mathrm{st}(e, \Delta_4^{i+1})$ (in this step we forget the colors on the vertices). If $v \in \mathrm{st}(e, \Delta_4^1)$ and $v' \in \mathrm{st}(e', \Delta_4^4)$, then since any path from v to v' must pass one vertex from the identified stars $\mathrm{st}(e', \Delta_4^i)$ $(i = 1, 2, 3)$, it follows that $\mathrm{dist}(v, v') \ge 4$. Hence by identifying $\mathrm{st}(e, \Delta_4^1)$ with $\mathrm{st}(e', \Delta_4^4)$ and removing e, e' in a flag handle addition, we obtain a new flag 3-manifold Γ with $\beta_1(\Gamma) = 1$. Both $\mathrm{lk}(e, \Delta_4^1)$ and $\mathrm{lk}(e', \Delta_4^4)$ are 4-cycles. Hence by Lemma 5.9,

$$\gamma_2(\Gamma) = \gamma_2(\Delta_{16}) + 2\gamma_1(\mathrm{lk}(e, \Delta_4^1)) + 16 = 16.$$

Step 3: Generate a flag 3-manifold with arbitrary β_1. This is done by taking the flag connected sum of b copies of Γ along the stars of edges whose links are 4-cycles. The resulting complex has $\gamma_2 = 16b$ and $\beta_1 = b$. □

Acknowledgements

The research was part of Lab of Geometry at Michigan projects offered by the Department of Mathematics at University of Michigan during the winter semester

of 2019. We would like to thank Harrison Bray and many others who coordinated the projects. We also thank the referee for valuable feedback.

References

[Alexander 1930] J. W. Alexander, "The combinatorial theory of complexes", *Ann. of Math.* (2) **31**:2 (1930), 292–320. MR Zbl

[Barnette 1973a] D. Barnette, "Graph theorems for manifolds", *Israel J. Math.* **16** (1973), 62–72. MR Zbl

[Barnette 1973b] D. Barnette, "A proof of the lower bound conjecture for convex polytopes", *Pacific J. Math.* **46** (1973), 349–354. MR Zbl

[Björner and Lutz 2000] A. Björner and F. H. Lutz, "Simplicial manifolds, bistellar flips and a 16-vertex triangulation of the Poincaré homology 3-sphere", *Experiment. Math.* **9**:2 (2000), 275–289. MR Zbl

[Charney and Davis 1995] R. Charney and M. Davis, "The Euler characteristic of a nonpositively curved, piecewise Euclidean manifold", *Pacific J. Math.* **171**:1 (1995), 117–137. MR Zbl

[Datta and Murai 2017] B. Datta and S. Murai, "On stacked triangulated manifolds", *Electron. J. Combin.* **24**:4 (2017), art. id. 4.12. MR Zbl

[Davis and Okun 2001] M. W. Davis and B. Okun, "Vanishing theorems and conjectures for the ℓ^2-homology of right-angled Coxeter groups", *Geom. Topol.* **5** (2001), 7–74. MR Zbl

[Eilenberg and Steenrod 1952] S. Eilenberg and N. Steenrod, *Foundations of algebraic topology*, Princeton University Press, 1952. MR Zbl

[Forman 2007] R. Forman, "Topics in combinatorial differential topology and geometry", pp. 133–205 in *Geometric combinatorics*, edited by E. Miller et al., IAS/Park City Math. Ser. **13**, Amer. Math. Soc., Providence, RI, 2007. MR Zbl

[Gal 2005] S. R. Gal, "Real root conjecture fails for five- and higher-dimensional spheres", *Discrete Comput. Geom.* **34**:2 (2005), 269–284. MR Zbl

[Gromov 1987] M. Gromov, "Hyperbolic groups", pp. 75–263 in *Essays in group theory*, edited by S. M. Gersten, Math. Sci. Res. Inst. Publ. **8**, Springer, 1987. MR Zbl

[Juhnke-Kubitzke et al. 2018] M. Juhnke-Kubitzke, S. Murai, I. Novik, and C. Sawaske, "A generalized lower bound theorem for balanced manifolds", *Math. Z.* **289**:3-4 (2018), 921–942. MR Zbl

[Kalai 1987] G. Kalai, "Rigidity and the lower bound theorem, I", *Invent. Math.* **88**:1 (1987), 125–151. MR Zbl

[Karu 2006] K. Karu, "The cd-index of fans and posets", *Compos. Math.* **142**:3 (2006), 701–718. MR Zbl

[Klee and Novik 2016] S. Klee and I. Novik, "Lower bound theorems and a generalized lower bound conjecture for balanced simplicial complexes", *Mathematika* **62**:2 (2016), 441–477. MR Zbl

[Lee 1997] C. W. Lee, "Subdivisions and triangulations of polytopes", pp. 271–290 in *Handbook of discrete and computational geometry*, edited by J. E. Goodman and J. O'Rourke, CRC, Boca Raton, FL, 1997. MR Zbl

[Lutz 2003] F. Lutz, "Triangulated Manifolds with Few Vertices: Geometric 3-Manifolds", preprint, 2003. arXiv

[Lutz 2011] F. H. Lutz, "The manifold page", webpage, 2011, https://page.math.tu-berlin.de/~lutz/stellar/.

[Lutz and Nevo 2016] F. H. Lutz and E. Nevo, "Stellar theory for flag complexes", *Math. Scand.* **118**:1 (2016), 70–82. MR Zbl

[Murai 2015] S. Murai, "Tight combinatorial manifolds and graded Betti numbers", *Collect. Math.* **66**:3 (2015), 367–386. MR Zbl

[Nevo 2007] E. Nevo, "Higher minors and Van Kampen's obstruction", *Math. Scand.* **101**:2 (2007), 161–176. MR Zbl

[Nevo and Petersen 2011] E. Nevo and T. K. Petersen, "On γ-vectors satisfying the Kruskal–Katona inequalities", *Discrete Comput. Geom.* **45**:3 (2011), 503–521. MR Zbl

[Novik and Swartz 2009] I. Novik and E. Swartz, "Socles of Buchsbaum modules, complexes and posets", *Adv. Math.* **222**:6 (2009), 2059–2084. MR Zbl

[Seifert and Threlfall 1980] H. Seifert and W. Threlfall, *Seifert and Threlfall: a textbook of topology*, Pure and Applied Mathematics **89**, Academic Press, New York, 1980. MR Zbl

[Sulanke and Lutz 2009] T. Sulanke and F. H. Lutz, "Isomorphism-free lexicographic enumeration of triangulated surfaces and 3-manifolds", *European J. Combin.* **30**:8 (2009), 1965–1979. MR Zbl

Received: 2020-03-31 Revised: 2020-07-20 Accepted: 2020-08-06

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

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

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

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

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

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

INVOLVE 13:4 (2020)
https://doi.org/10.2140/involve.2020.13.705

Some new Gompertz fractional difference equations

Tom Cuchta and Brooke Fincham

(Communicated by Martin J. Bohner)

We introduce three new fractional Gompertz difference equations using the Riemann–Liouville discrete fractional calculus. These three models are based a nonfractional Gompertz difference equation, and they differ depending on whether a fractional operator replaces the difference operator, the integral operator defining the logarithm, or both simultaneously. An explicit solution to one of them is achieved with restricted parameters and recurrence relation solutions are derived for all three. Finally, we fit these models to data to compare them with a previously published discrete fractional Gompertz model and the continuous model.

1. Introduction

The Gompertz function dates back to the study of human mortality conducted by Benjamin Gompertz [1825]. Since then, it has been used in numerous ways, particularly as a population growth model; e.g., see [Alves et al. 2019; Easton 1999; Jane et al. 2020; Laird 1964; Pezzini et al. 2019; Winsor 1932]. The Gompertz differential equation is

$$y' = -ry \ln\left(\frac{y}{K}\right),$$

which has closed-form general solution as the iterated exponential $y(t) = Ke^{Ce^{-rt}}$.

In contrast to the continuous Gompertz model, we are concerned with a discrete theory. The fundamental operator of the "forward" difference calculus is the Δ operator that obeys $\Delta f(t) = f(t+1) - f(t)$. An enormous amount of research in discrete analogues to the theory of differential equations has been conducted for these operators, and there are many well-known popular introductions to it such as [Bohner and Peterson 2001; Elaydi 2005; Kelley and Peterson 2010]. There are numerous difference equations that have been called a "Gompertz difference equation" in the

MSC2020: primary 39A60; secondary 26A33, 92D25, 33E12.
Keywords: discrete fractional calculus, Gompertz, parameter estimation, data fitting.

literature, such as in [Akın et al. 2020; Nobile et al. 1982; Satoh 2000]. We now consider the Δ-Gompertz model from [Bohner 2005, (12)], defined by

$$\Delta y = (\ominus r)(a + \tilde{L}_y)y, \quad y(t_0) = y_0, \tag{1}$$

where r and a are constants and

$$\tilde{L}_y(t, t_0) = \sum_{k=t_0}^{t-1} \frac{y(k+1) - y(k)}{y(k)} \tag{2}$$

is the Δ-logarithm, originally defined in [Bohner 2005, (3)]. The model (1) originally used time-dependent r and a, but we will not consider that generality here.

We will generalize (1) to a model using fractional sum and difference operators. Fractional differential calculus dates back to the early days of the theory of calculus [Ross 1975]. The fundamental idea of fractional calculus is the extension of derivatives and integrals to noninteger order. The fractional discrete calculus dates back to [Kuttner 1957], with an explosion of interest in more recent times [Abdeljawad 2011; Atici and Eloe 2009; Dzieliński et al. 2010; Ferreira 2013; Miller and Ross 1989; Wu et al. 2015]; see in particular the recent monograph [Goodrich and Peterson 2015]. An analogue of the Gompertz differential equation using fractional differential operators has also been developed and studied [Solís-Pérez et al. 2019; Frunzo et al. 2019]. Fractional Gompertz difference equations are not a new idea; for example, see [Wang et al. 2014; Bolton et al. 2015; Atıcı and Şengül 2010], which use various fractional difference operators. We will take one model from the literature for comparison to the models we will propose: Atıcı et al. defined a fractional Δ-difference equation in [Atıcı et al. 2017, (3)] by

$$(c - b\log y(t))y(t) = \begin{cases} \Delta^{\nu} y^{\sigma}(t-\nu), & \nu \in (0, 1.5) \setminus \{1\}, \\ \Delta y(t), & \nu = 1, \end{cases} \tag{3}$$

where Δ^{ν} denotes the discrete Riemann–Liouville Δ-fractional difference operator. We emphasize that (3) uses the classical logarithm and not a logarithm resembling (2). This is the fundamental change we have made by choosing (1) as the basis for our models. The equation (3) has the following form, convenient for numerical computation [Atıcı et al. 2017, p. 323]:

$$y(t+1) = \begin{cases} (c - b\log(y(t)))y(t) - \sum_{s=0}^{t-1} \frac{\Gamma(t-\nu-s)}{\Gamma(t-s+1)\Gamma(-\nu)} y(s+1), & \nu \in (0, 1.5) \setminus \{1\}, \\ (c + 1 - b\log(y(t)))y(t), & \nu = 1. \end{cases} \tag{4}$$

The recurrence (4) was used to fit the model (3) to data collected on mice tumors.

2. Preliminaries and definitions

For $a, b \in \mathbb{R}$ with $a - b$ a positive integer, we define as in [Goodrich and Peterson 2015, p. 1], henceforth abbreviated [GP15], the sets

$$\mathbb{N}_a := \{a, a+1, a+2, a+3, \dots\}, \quad \mathbb{N}_a^b := \{a, a+1, a+2, a+3, \dots, b\}.$$

Throughout, we use the usual convention for empty products that $\prod_{j=a}^{b} f(j) = 1$ and for empty sums $\sum_{j=a}^{b} f(j) = 0$ whenever $b < a$. The backwards jump operator $\rho : \mathbb{N}_a \to \mathbb{N}_{a-1}$ is given by [GP15, p. 150], where $\rho(t) = t - 1$ for $t > 0$ and $\rho(0) = 0$. The backwards difference operator is [GP15, p. 150] $\nabla y(t) = y(t) - y(\rho(t))$. If $f : \mathbb{N}_{a+1} \to \mathbb{R}$, $a \in \mathbb{N}_a$, and $b \in \mathbb{N}_{a+1}$, then the discrete ∇-integral of f is the summation defined by

$$\int_a^b f(t)\, \nabla t = \sum_{t=a+1}^{b} f(t)$$

[GP15, Definition 3.31].

The gamma function $\Gamma : (0, \infty) \to [1, \infty)$ is defined by

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x}\, \mathrm{d}x$$

[Artin 1964, (2.1)]. Integration by parts shows that the functional equation

$$\Gamma(t+1) = t\Gamma(t)$$

holds [Artin 1964, (2.2)], which extends the domain of Γ to $\mathbb{R}\backslash\{0, -1, -2, \dots\}$. For $r \in \mathbb{Z}^+$ and $t \in \mathbb{N}_0$, the rising factorial function is

$$t^{\bar{r}} = \frac{\Gamma(t+r)}{\Gamma(t)}$$

[GP15, Definition 3.3]. The gamma function is primarily used here to define the ∇-Taylor monomials, H_μ, which are defined for $\mu \neq -1, -2, -3, \dots$ and $t \in \mathbb{N}_a$ by [GP15, Definition 3.56],

$$H_\mu(t, a) = \frac{(t-a)^{\bar{\mu}}}{\Gamma(\mu+1)} = \frac{\Gamma(t-a+\mu)}{\Gamma(\mu+1)\Gamma(t-a)},$$

so if $\mu = 0$, then we observe $H_0(t, a) = 1$. In particular, we observe the following useful identity that we will occasionally use:

$$H_\mu(t, t-1) = \frac{\Gamma(t-(t-1)+\mu)}{\Gamma(\mu+1)\Gamma(t-(t-1))} = \frac{\Gamma(\mu+1)}{\Gamma(\mu+1)\Gamma(1)} = 1.$$

Given $p : \mathbb{N}_a \to \mathbb{R}$ such that $1 - p(t) \neq 0$ for all t, we define the $\boxminus$ operator by

$$(\boxminus p)(t) = \frac{-p(t)}{1 - p(t)}$$

[GP15, (3.10)]. In particular, if $p(t) = r$ is a constant function, then

$$(\boxminus r)(t) = \frac{-r}{1-r}$$

is also a constant function. For $p : \mathbb{N}_{a+1} \to \mathbb{R}$ obeying $1 - p(t) \neq 0$, the ∇-exponential function E_p is given by [GP15, (3.7)]

$$E_p(t,s) = \begin{cases} \prod_{\tau=s+1}^{t} 1/(1-p(\tau)), & t \in \mathbb{N}_s, \\ \prod_{\tau=t+1}^{s} [1-p(\tau)], & t \in \mathbb{N}_a^{s-1}. \end{cases} \tag{5}$$

The following result confirms that the ∇-exponential solves a first-order difference equation.

Theorem 1 [GP15, Theorem 3.6]. *Assume $p : \mathbb{N}_{a+1} \to \mathbb{R}$ such that $1 - p(t) \neq 0$, and let $t \in \mathbb{N}_{a+1}$. The ∇-exponential $y(t) = E_p(t,s)$ is the unique solution of the initial value problem*

$$\nabla y(t) = p(t) y(t), \quad y(s) = 1.$$

If $f : \mathbb{N}_{a+1} \to \mathbb{R}$, $\nu > 0$, and $t \in \mathbb{N}_a$, then the ν-th order ∇-fractional sum of f at t is defined by [GP15, Definition 3.58]

$$\nabla_a^{-\nu} f(t) = \int_a^t H_{\nu-1}(t, \rho(s)) f(s) \, \nabla s,$$

where we let $\nabla_a^{-0} f(t) = f(t)$. Let $f : \mathbb{N}_a \to \mathbb{R}$, $\nu > 0$, $t \in \mathbb{N}_{a+1}$, and $N \in \mathbb{N}_1$, where $N - 1 < \nu \leq N$. The ν-th order ∇-fractional difference of f at t is similarly defined by [GP15, (3.32)]

$$\nabla_a^{\nu} f(t) = \int_a^t H_{-\nu-1}(t, \rho(s)) f(s) \, \nabla s. \tag{6}$$

From the definition, it is easy to see that $\nabla_a^{\nu} f(a+1) = f(a+1)$. The following result shows how fractional differences and summations compose.

Theorem 2 [GP15, Theorem 3.109]. *If $f : \mathbb{N}_a \to \mathbb{R}$ and $\nu, \mu > 0$, then*

$$\nabla_a^{\nu} \nabla_a^{-\mu} f(t) = \nabla_a^{\nu-\mu} f(t).$$

The fractional difference operator subtracts from the order of the ∇-Taylor monomials as $\nabla_{t_0}^{\nu} H_{\mu}(t, t_0) = H_{\mu-\nu}(t, t_0)$ [GP15, Theorem 3.93(ii)], and so for any constant C

$$\nabla_a^{\nu} C = C \nabla_a^{\nu} H_0(t, a) = C H_{-\nu}(t, a).$$

In order to solve fractional initial value problems in closed form, the ∇-convolution is needed. Let $f, g : \mathbb{N}_{a+1} \to \mathbb{R}$, and define the ∇-convolution by [GP15, Definition 3.77]

$$(f * g)(t) = \int_a^t f(t - \rho(\tau) + a) g(\tau) \, \nabla\tau.$$

The ∇-Mittag-Leffler function, $E_{p,\alpha,\beta}$, where $|p| < 1,\ \alpha > 0,\ \beta \in \mathbb{R}$, and $t \in \mathbb{N}_a$, is defined by [GP15, Definition 3.98]

$$E_{p,\alpha,\beta}(t, a) = \sum_{k=0}^{\infty} p^k H_{\alpha k+\beta}(t, a).$$

The following result is a discrete fractional analogue of the variation of constants formula for first-order initial value problems.

Theorem 3 [GP15, Theorem 3.104]. *If $f : \mathbb{N}_a \to \mathbb{R},\ |c| < 1$, and $0 < \nu < 1$, then the unique solution of the fractional initial value problem*

$$\begin{cases} \nabla^{\nu}_{\rho(a)} x(t) + c x(t) = f(t), & t \in \mathbb{N}_{a+1}, \\ x(a) = A, & A \in \mathbb{R}, \end{cases}$$

is given by

$$x(t) = [E_{-c,\nu,\nu-1}(\cdot, \rho(a)) * f(\cdot)](t) + [A(c+1) - f(a)] E_{-c,\nu,\nu-1}(t, \rho(a)). \quad (7)$$

The recurrence (4) came from applying a Δ-analogue [Atıcı and Şengül 2010, p. 322] of the following lemma to (3).

Lemma 4. *If $f : \mathbb{N}_{s+1} \to \mathbb{R}$, then for $t \in \mathbb{N}_{s+1}$*

$$\nabla^{\nu}_s f(t) = f(t) + \sum_{k=s+1}^{t-1} \frac{\Gamma(t - k - \nu) f(k)}{\Gamma(-\nu)\Gamma(t - k + 1)}.$$

Proof. Calculating directly from the definition,

$$\begin{aligned} \nabla^{\nu}_s f(t) &= \int_s^t H_{-\nu-1}(t, \rho(\tau)) f(\tau) \, \nabla\tau = \sum_{k=s+1}^{t} H_{-\nu-1}(t, \rho(k)) f(k) \\ &= H_{-\nu-1}(t, \rho(t)) f(t) + \sum_{k=s+1}^{t-1} H_{-\nu-1}(t, \rho(k)) f(k) \\ &= f(t) + \sum_{k=s+1}^{t-1} \frac{\Gamma(t - k - \nu) f(k)}{\Gamma(-\nu)\Gamma(t - k + 1)}, \end{aligned}$$

completing the proof. □

We now define the various ∇-logarithms we will use. The natural ∇-analogue of (2) is

$$\mathrm{L}_y(t, t_0) = \int_{t_0}^{t} \frac{y^{\nabla}(\tau)}{y(\tau)} \nabla\tau, \quad t \in \mathbb{N}_{t_0+1}.$$

We define the first fractional ∇_ν-logarithm $\mathrm{L}_{y,\nu}$ by

$$\mathrm{L}_{y,\nu}(t, t_0) = \left(\nabla_{t_0}^{-\nu} \frac{y^{\nabla}}{y}\right)(t), \quad t \in \mathbb{N}_{t_0+1}. \tag{8}$$

we define the second fractional ∇_ν-logarithm $\ell_{y,\nu}$ by

$$\ell_{y,\nu}(t, t_0) = \int_{t_0}^{t} \frac{\nabla_{t_0}^{\nu} y(\tau)}{y(\tau)} \nabla\tau, \quad t \in \mathbb{N}_{t_0+1}, \tag{9}$$

and the third ∇_ν-logarithm $\Lambda_{y,\nu}$ by

$$\Lambda_{y,\nu}(t, t_0) = \left(\nabla_{t_0}^{-\nu} \frac{\nabla_{t_0}^{\nu} y}{y}\right)(t), \quad t \in \mathbb{N}_{t_0+1}. \tag{10}$$

3. Three Gompertz fractional ∇-difference equations

The natural ∇-analogue of (1) is

$$\nabla y = (\boxminus r)(a + \mathrm{L}_y) y, \quad y(t_0) = y_0, \tag{11}$$

whose unique solution is easily found using the same method used in [Cuchta and Streipert 2020]. There are many ways in which a ∇-fractional analogue of (11) could be defined, depending on which operators in the equation are made fractional. We will consider three such analogues in this article. In light of (3), we will allow the order of our fractional derivative to lie in the interval $(0, 1.5)$, where $\nu = 1$ corresponds to the nonfractional ∇ difference. First, we preserve the nonfractional difference on y from (11), but we use the fractional ∇-logarithm (8) to obtain

$$\nabla y = (\boxminus r) y (a + \mathrm{L}_{y,\nu}), \quad y(t_0) = y_0. \tag{12}$$

If instead the difference on y in (11) is a fractional difference, but the ∇-logarithm (9) is used, then we obtain

$$\nabla_{t_0}^{\nu} y(t) = (\boxminus r) y (a + \ell_{y,\nu}), \quad y(t_0 + 1) = y_0. \tag{13}$$

Finally, if we make the difference on y in (11) a fractional difference and use the fractional ∇-logarithm (10), then we obtain

$$\nabla_{t_0}^{\nu} y(t) = (\boxminus r) y (a + \Lambda_{y,\nu}), \quad y(t_0 + 1) = y_0. \tag{14}$$

Note that all these logarithms reduce to L_y as $\nu \to 1$. We now find a closed-form solution of (12) in terms of the ∇-Mittag-Leffler function.

Theorem 5. *If* $0 < \nu < 1$, $r < \frac{1}{2}$, *and, for* $t \in \mathbb{N}_{t_0}$,

$$\sum_{k=t_0+1}^{t} E_{\boxminus r,\nu,\nu-1}(t-k+1+t_0, t_0) H_{-\nu}(k, t_0) \neq -1 - \frac{1-r}{ar}, \tag{15}$$

then the initial value problem (12) *has the unique solution* $y(t) = y_0 E_p(t, t_0)$ *for* $t \in \mathbb{N}_{t_0}$, *where*

$$p(t) = (\boxminus r)a + (\boxminus r)\big(E_{\boxminus r,\nu,\nu-1}(\cdot, t_0) * aH_{-\nu}(\cdot, t_0)\big)(t).$$

Proof. If $z(t) = a + \mathrm{L}_{y,\nu}(t, t_0)$, then

$$(\nabla_{t_0}^{\nu} z)(t) = aH_{-\nu}(t, t_0) + \frac{\nabla y(t)}{y(t)}. \tag{16}$$

Now compute

$$\begin{aligned} a + \mathrm{L}_{y,\nu}(t_0+1, t_0) &= a + \left(\nabla_{t_0}^{-\nu} \frac{\nabla y}{y}\right)(t_0+1) \\ &= a + \int_{t_0}^{t_0+1} H_{\nu-1}(t_0+1, \rho(s)) \frac{\nabla y(s)}{y(s)} \nabla s \\ &= a + H_{\nu-1}(t_0+1, t_0) \frac{\nabla y(t_0+1)}{y(t_0+1)} = a + 1 - \frac{y_0}{y(t_0+1)}. \end{aligned}$$

We expand the difference equation (12) at $t = t_0 + 1$ to get

$$y(t_0+1) - y_0 = -\frac{r}{1-r} y(t_0+1)\left(a + 1 - \frac{y_0}{y(t_0+1)}\right),$$

and hence

$$y(t_0+1) = \frac{y_0}{1+ra}.$$

Immediately, we obtain $z(t_0+1) = a + \mathrm{L}_{y,\nu}(t_0+1, t_0) = a(1-r)$. Using the definition of z, (16), and dividing the difference equation (12) by y, the following initial value problem for z is derived:

$$\nabla_{t_0}^{\nu} z + (-\boxminus r) z = aH_{-\nu}(t, t_0), \quad z(t_0+1) = a(1-r). \tag{17}$$

Theorem 3 implies that (17) has the unique solution

$$z(t) = a + \big(E_{\boxminus r,\nu,\nu-1}(\cdot, t_0) * aH_{-\nu}(\cdot, t_0)\big)(t).$$

Therefore, we have the initial value problem $\nabla y = (\boxminus r) z y$, $y(t_0) = y_0$. We need to ensure that the function $h : \mathbb{N}_{t_0} \to \mathbb{R}$ given by $h(t) = (\boxminus r) z(t)$ obeys $h(t) \neq 1$. Expanding out the definition of h reveals

$$1 \neq \left(\frac{-r}{1-r}\right)\left(a + ak \sum_{k=t_0}^{t} E_{\boxminus r,\nu,\nu-1}(t-k+1+t_0, t_0) H_{-\nu}(k, t_0)\right),$$

and by algebraic rearrangement, this becomes the condition (15). Thus, we may appeal to Theorem 1 to complete the proof. □

Theorem 5 provides an impractical solution for numerical computation for a variety of reasons. In particular, the ∇-Mittag-Leffler function requires the evaluation of an infinite series, the order is restricted to $0 < \nu < 1$, and condition (15) is not trivial to verify. So, we now derive a recurrence relation for numerical computation of the solution of (12).

Theorem 6. *The initial value problem* (12) *is algebraically equivalent to the recurrence*

$$y(t) = y_0 \prod_{k=t_0+1}^{t} \frac{1}{1-(\boxminus r)(a+z(k))}, \quad t \in \mathbb{N}_{t_0},$$

where

$$z(t) = -ra - (1-r) \sum_{k=t_0+1}^{t-1} \frac{\Gamma(t-k-\nu)z(k)}{\Gamma(-\nu)\Gamma(t-k+1)}, \quad t \in \mathbb{N}_{t_0+1}. \tag{18}$$

Proof. If $z(t) = \mathrm{L}_{y,\nu}(t, t_0)$, then

$$\nabla_{t_0}^{\nu} z(t) = \frac{y^{\nabla}(t)}{y(t)} = (\boxminus r)(a+z(t)).$$

Apply Lemma 4 to the left-hand side to get

$$z(t) + \sum_{k=t_0+1}^{t-1} \frac{\Gamma(t-k-\nu)z(k)}{\Gamma(-\nu)\Gamma(t-k+1)} = (\boxminus r)(a+z(t)). \tag{19}$$

We obtain (18) by rearranging (19) and solving for $z(t)$. Substituting $z(t)$ back into (12), we observe

$$\nabla y(t) = (\boxminus r)(a+z(t))y(t), \quad y(t_0) = y_0.$$

By Theorem 1, we have $y(t) = y_0 E_{(\boxminus r)(a+z(t))}(t, t_0)$, and applying (5) completes the proof. □

Theorem 7. *The initial value problem* (13) *is algebraically equivalent to the following recurrence for* $t > t_0 + 1$:

$$y(t) = \frac{-\sum_{k=t_0+1}^{t-1} H_{-\nu-1}(t, k-1)y(k)}{1-r+r\left(a+1+\sum_{k=t_0+1}^{t-1} \frac{1}{y(k)} \sum_{j=t_0+1}^{k} H_{-\nu-1}(k, j-1)y(j)\right)}, \quad t \in \mathbb{N}_{t_0+2}.$$

Proof. Expand the fractional derivative and the logarithm in (13) to obtain

$$\int_{t_0}^{t} H_{-\nu-1}(t, \rho(s))y(s)\,\nabla s = (\boxminus r)y(t)\left(a + \int_{t_0}^{t} \frac{\nabla_{t_0}^{\nu} y(\tau)}{y(\tau)}\,\nabla\tau\right).$$

Separating the final term of the ∇-integrals and rearranging algebraically yields

$$y(t) = \frac{((\boxminus r) - 1)\int_{t_0}^{t-1} H_{-\nu-1}(t, \rho(s))y(s)\,\nabla s}{1 - (\boxminus r)\big(a + 1 + \int_{t_0}^{t-1} \frac{\nabla_{t_0}^{\nu} y(\tau)}{y(\tau)}\,\nabla\tau\big)}$$
$$= \frac{((\boxminus r) - 1)\int_{t_0}^{t-1} H_{-\nu-1}(t, \rho(s))y(s)\,\nabla s}{1 - (\boxminus r)\big(a + 1 + \int_{t_0}^{t-1} \frac{1}{y(\tau)}\int_{t_0}^{\tau} H_{-\nu-1}(\tau, \rho(s))y(s)\,\nabla s\,\nabla\tau\big)},$$

and writing the ∇-integrals as summations, replacing $(\boxminus r)$ with $-r/(1-r)$, and routine algebraic simplification completes the proof. □

Theorem 8. *The initial value problem* (14) *is algebraically equivalent to the recurrence*

$$y(t) = \frac{-\sum_{k=t_0+1}^{t-1} H_{-\nu-1}(t, k-1)y(k)}{1 - r + r\big(a + 1 + \sum_{k=t_0+1}^{t-1} \frac{H_{\nu-1}(t,k-1)}{y(k)} \sum_{j=t_0+1}^{k} H_{-\nu-1}(k, j-1)y(j)\big)}, \quad t \in \mathbb{N}_{t_0+2}. \tag{20}$$

Proof. Expand (14) using definitions to get

$$\int_{t_0}^{t} H_{-\nu-1}(t, \rho(\tau))y(\tau)\,\nabla\tau = (\boxminus r)y(t)\bigg(a + \int_{t_0}^{t} H_{\nu-1}(t, \rho(\tau))\frac{\nabla_{t_0}^{\nu} y(\tau)}{y(\tau)}\,\nabla\tau\bigg).$$

Separate the final terms from the ∇-integrals and rearrange algebraically to obtain

$$y(t) = \frac{((\boxminus r) - 1)\int_{t_0}^{t-1} H_{-\nu-1}(t, \rho(\tau))y(\tau)\,\nabla\tau}{1 - (\boxminus r)a - (\boxminus r) - (\boxminus r)\int_{t_0}^{t-1} H_{\nu-1}(t, \rho(\tau))\frac{\nabla_{t_0}^{\nu} y(\tau)}{y(\tau)}\,\nabla\tau}$$
$$= \frac{((\boxminus r) - 1)\int_{t_0}^{t-1} H_{-\nu-1}(t, \rho(\tau))y(\tau)\,\nabla\tau}{1 - (\boxminus r)\big(a + 1 + \int_{t_0}^{t-1} \frac{H_{\nu-1}(t,\rho(\tau))}{y(\tau)}\int_{t_0}^{\tau} H_{-\nu-1}(\tau, \rho(s))y(s)\,\nabla s\,\nabla\tau\big)},$$

and writing the ∇-integrals as summations and applying the definition of $(\boxminus r)$ completes the proof. □

4. Curve fitting and comparison of models

We will now present curve fittings of our three models (12), (13), and (14), the model (3), and the continuous model to data that had previously fit with (continuous) Gompertz curves. In order to test the fit of our models, we have programmed[1] each model into the Python programming language and used the `curve_fit` function from `scipy.optimize`. According to its documentation, the `curve_fit` function ultimately implements the Levenberg–Marquardt algorithm for solving nonlinear least squares problems. To implement the solutions of our models, we used the

[1] https://github.com/tomcuchta/cuchtafinchamdiscretefractionalgompertz

	v	a	r	y_0	RSS	SE
(12)	$5.58\cdot10^{-1}$	$-2.38\cdot10^{-1}$	$2.33\cdot10^{0}$	$1.78\cdot10^{2}$	$8.49\cdot10^{6}$	$2.43\cdot10^{2}$
(13)	$1.13\cdot10^{0}$	$-1.52\cdot10^{0}$	$4.89\cdot10^{-1}$	$2.07\cdot10^{2}$	$8.61\cdot10^{6}$	$2.44\cdot10^{2}$
(14)	$8.77\cdot10^{-1}$	$-1.73\cdot10^{0}$	$4.87\cdot10^{-1}$	$2.08\cdot10^{2}$	$8.60\cdot10^{6}$	$2.44\cdot10^{2}$
	v	c	b	y_0	RSS	SE
(3)	$5.88\cdot10^{-1}$	$4.46\cdot10^{0}$	$6.79\cdot10^{-1}$	$7.45\cdot10^{1}$	$8.61\cdot10^{6}$	$2.44\cdot10^{2}$
	L_∞	G	t_0		RSS	SE
ctn	$5.85\cdot10^{2}$	$3.81\cdot10^{-1}$	$1.79\cdot10^{0}$		$8.56\cdot10^{6}$	$2.43\cdot10^{2}$

Table 1. Results of curve fittings to data from [Hilling et al. 2016b] and their goodness of fit. The continuous curve was given there by the three-parameter formula $L_\infty e^{-e^{-G(t-t_0)}}$.

recurrences in Theorems 6, 7, and 8 directly. The initial condition argument t_0 was always chosen so that the data point with smallest independent variable agreed with the minimum of the domain of the solution. The effect this has is to horizontally shift the solutions so that they all "begin" at the same value of the independent variable in a given data set. Because those recurrence relations are self-referential at every iteration, they are computationally inefficient without optimizing the code in some way. So, we used the technique of memoization, meaning that we stored all newly generated values for a solution in a dictionary data structure to reference when computing the next value of the solution.

We present the parameters found, and as done in [Atıcı et al. 2017], we computed the residual sum of squared errors (RSS) and the standard error (SE). The RSS is defined by

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - Y(t_i))^2,$$

where n denotes the number of data points, (t_i, y_i) denotes the data points themselves, and $Y(t_i)$ denotes the prediction of the fitted model at t_i, and SE is given by

$$\mathrm{SE} = \sqrt{\frac{\mathrm{RSS}}{n-k}},$$

where n denotes the total number of data points and k is the number of parameters of the model.

The article [Hilling et al. 2016b] concerns the relationship between the age (in years) and length (in millimeters) of 155 *ictalurus punctatus* ("channel catfish") caught in Cheat Lake in West Virginia. The lengths of the fish were measured, and a technique to determine their age was performed. Their data set is publicly accessible

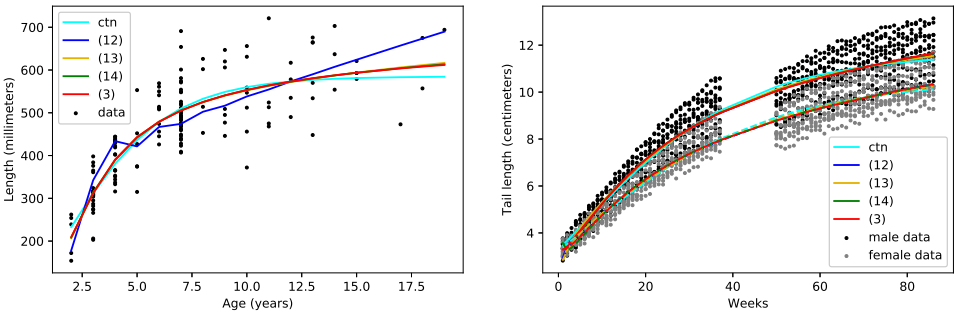


Figure 1. Left: Optimal models for the data from [Hilling et al. 2016b]. Right: Optimal models for the data from [Yang et al. 2019b]. Male (solid curve) and female (dashed curve) tail-length data. The curve for (12) is difficult to see because it is below the other curves.

male tail-length data						
	v	a	r	y_0	RSS	SE
(12)	$8.92\cdot10^{-1}$	$-8.34\cdot10^{-1}$	$9.55\cdot10^{-2}$	$2.94\cdot10^{0}$	$2.72\cdot10^{4}$	$5.33\cdot10^{0}$
(13)	$1.13\cdot10^{0}$	$-1.53\cdot10^{0}$	$5.79\cdot10^{-2}$	$2.85\cdot10^{0}$	$2.72\cdot10^{4}$	$5.33\cdot10^{0}$
(14)	$8.25\cdot10^{-1}$	$-1.87\cdot10^{0}$	$1.39\cdot10^{-1}$	$3.56\cdot10^{0}$	$2.72\cdot10^{4}$	$5.33\cdot10^{0}$
	v	c	b	y_0	RSS	SE
(3)	$2.30\cdot10^{-2}$	$1.15\cdot10^{0}$	$1.02\cdot10^{-1}$	$3.07\cdot10^{0}$	$2.72\cdot10^{4}$	$5.33\cdot10^{0}$
	A	B	K		RSS	SE
ctn	$1.17\cdot10^{1}$	$1.27\cdot10^{0}$	$4.52\cdot10^{-2}$		$2.72\cdot10^{4}$	$5.33\cdot10^{0}$

female tail-length data						
	v	a	r	y_0	RSS	SE
(12)	$8.60\cdot10^{-1}$	$-6.76\cdot10^{-1}$	$1.03\cdot10^{-1}$	$2.89\cdot10^{0}$	$1.54\cdot10^{4}$	$4.35\cdot10^{0}$
(13)	$1.12\cdot10^{0}$	$-1.51\cdot10^{0}$	$4.62\cdot10^{-2}$	$2.84\cdot10^{0}$	$1.54\cdot10^{4}$	$4.35\cdot10^{0}$
(14)	$7.56\cdot10^{-1}$	$-1.72\cdot10^{0}$	$1.82\cdot10^{-1}$	$3.56\cdot10^{0}$	$1.54\cdot10^{4}$	$4.36\cdot10^{0}$
	v	c	b	y_0	RSS	SE
(3)	$3.58\cdot10^{-2}$	$1.14\cdot10^{0}$	$1.23\cdot10^{-1}$	$3.35\cdot10^{0}$	$1.54\cdot10^{4}$	$4.35\cdot10^{0}$
	A	B	K		RSS	SE
ctn	$1.05\cdot10^{1}$	$1.18\cdot10^{0}$	$3.95\cdot10^{-2}$		$1.53\cdot10^{4}$	$4.35\cdot10^{0}$

Table 2. Results of curve fittings to data from [Yang et al. 2019b] and their goodness of fit. The continuous curve was given there by the three-parameter formula $Ae^{-Be^{-Kt}}$.

[Hilling et al. 2016a] and contains columns for the age and length (both positive integers), which we used without modification. Their analysis of the various growth curves concluded that the von Bertalanffy growth (proportional to $1 - e^{-k(t-t_0)}$, commonly used in length-based methods for fish population estimates [Pauly and Morgan 1987]) was the model of best fit among those considered. The results of our work are summarized in Table 1 and visualized in Figure 1, left. We see that the five models have essentially identical RSS, indicating that they all fit equally well. The models (12), (14), and (3) have optimal ν parameter between 0.55 and 0.87, while the optimal ν for (13) is larger than 1. The fit of (12) is not monotone, while the other four models are visually quite similar.

In [Yang et al. 2019b], the logistic, Gompertz, and von Bertalanffy growth curves were fit to data collected about *plestiodon elegans* (a type of skink) pertaining to some of their physical traits — specifically, the tail length of thirteen male and eleven female individuals were measured once a week for 85 weeks. Their data set is publicly accessible [Yang et al. 2019a] and contains columns for the animal identification number, week number (both positive integers), and tail length in centimeters (accurate to two decimal places), which we used without modification. The study concluded that the continuous Gompertz model was the best fit for both sexes for tail length data. We have fit the models to male and female tail-length data, and our results are summarized in Table 2, and visualized in Figure 1, right. We observe that the RSS and SE are nearly identical for all curve fittings, indicating that the models fit similarly well. The models (12) and (14) have ν-value between 0.74 and 0.89, (13) always has a ν value larger than 1, and the model (3) always has small ν.

5. Conclusion

We have investigated three new natural fractional analogues of the Gompertz ∇-difference equation. We have found a closed form solution of one of them in terms of the ∇-Mittag-Leffler function, but with necessarily restricted parameters. We derived recurrence relations for all three. We have fit the models to data sets and compared the fit to the existing fractional Gompertz model by Atıcı et al. It appears that both of the data sets were fit nearly equally well by all four models according to RSS and SE, and so using biological considerations or more advanced statistical techniques would be necessary to decide which model is best. We observed that the four discrete fractional models under consideration did not perform better than their continuous counterparts in terms of RSS or SE.

There are numerous directions of future research involving these models. Finding simple sufficient conditions for the regressivity of the solution of the first new fractional model is of great interest, not only for the fractional model but also for the original nonfractional model. Many qualitative properties remain unexplored,

including analysis of stability and oscillation of their solutions. One interesting fractional generalization that we did not consider is taking the orders of the fractional difference and fractional sum in the third new fractional logarithm to be different. Such a model would gain an extra parameter, and so it may fit better in terms of RSS and SE. When comparing models with different numbers of parameters, a more robust statistical analysis should be conducted, e.g., the Akaike information criterion method.

Our choice of fractional derivative had some consequences that complicated some aspects of the work. Most importantly, in the proof of Theorem 5, the fractional derivative of $z(t) = a + \mathrm{L}_{y,\nu}(t, t_0)$ was computed yielding

$$(\nabla_{t_0}^{\nu} z)(t) = a H_{-\nu}(t, t_0) + \frac{\nabla y(t)}{y(t)}.$$

Since we used the Riemann–Liouville fractional difference, the derivative of the constant a was not zero, and we consequently had to use a convolution via Theorem 3 to express the closed-form solution of the first new model. This raises the general question of defining fractional Gompertz equations using the Caputo fractional differences (and others), where the fractional derivative of a constant is zero. Such models may be easier to analyze analytically, but may have other unforeseen downsides.

Acknowledgments

This research was made possible by NASA West Virginia Space Grant Consortium, Training Grant #NNX15AI01H. We also thank the referees for their valuable recommendations that improved the quality of this manuscript.

References

[Abdeljawad 2011] T. Abdeljawad, "On Riemann and Caputo fractional differences", *Comput. Math. Appl.* **62**:3 (2011), 1602–1611. MR Zbl

[Akın et al. 2020] E. Akın, N. Nesliye Pelen, I. Uğur Tiryaki, and F. Yalcin, "Parameter identification for Gompertz and logistic dynamic equations", *PLoS ONE* **15**:4 (2020), art. id. e0230582.

[Alves et al. 2019] W. J. Alves, E. B. Malheiros, N. K. Sakomura, E. P. da Silva, G. da Silva Viana, C. A. G. M. de Paula Reis, and R. M. Suzuki, "*In vivo* description of body growth and chemical components of egg-laying pullets", *Livestock Sci.* **220** (2019), 221–229.

[Artin 1964] E. Artin, *The gamma function*, Holt, Rinehart and Winston, New York, 1964. MR Zbl

[Atıcı and Şengül 2010] F. M. Atıcı and S. Şengül, "Modeling with fractional difference equations", *J. Math. Anal. Appl.* **369**:1 (2010), 1–9. MR Zbl

[Atici and Eloe 2009] F. M. Atici and P. W. Eloe, "Initial value problems in discrete fractional calculus", *Proc. Amer. Math. Soc.* **137**:3 (2009), 981–989. MR Zbl

[Atıcı et al. 2017] F. M. Atıcı, M. Atıcı, M. Belcher, and D. Marshall, "A new approach for modeling with discrete fractional equations", *Fund. Inform.* **151**:1-4 (2017), 313–324. MR

[Bohner 2005] M. Bohner, "The logarithm on time scales", *J. Difference Equ. Appl.* **11**:15 (2005), 1305–1306. MR Zbl

[Bohner and Peterson 2001] M. Bohner and A. Peterson, *Dynamic equations on time scales: an introduction with applications*, Birkhäuser, Boston, MA, 2001. MR Zbl

[Bolton et al. 2015] L. Bolton, A. H. J. J. Cloot, S. W. Schoombie, and J. P. Slabbert, "A proposed fractional-order Gompertz model and its application to tumour growth data", *Math. Med. Biol.* **32**:2 (2015), 187–207. MR Zbl

[Cuchta and Streipert 2020] T. Cuchta and S. Streipert, "Dynamic Gompertz model", *Appl. Math. Inf. Sci.* **14**:1 (2020), 9–17. MR

[Dzieliński et al. 2010] A. Dzieliński, D. Sierociuk, and G. Sarwas, "Some applications of fractional order calculus", *Bull. Polish Acad. Sci. Tech. Sci.* **58**:4 (2010), 583–592. Zbl

[Easton 1999] D. M. Easton, "X-ray survival as Gompertz growth in number killed", *J. Theoret. Biol.* **196**:1 (1999), 1–8.

[Elaydi 2005] S. Elaydi, *An introduction to difference equations*, 3rd ed., Springer, 2005. MR Zbl

[Ferreira 2013] R. A. C. Ferreira, "Existence and uniqueness of solution to some discrete fractional boundary value problems of order less than one", *J. Difference Equ. Appl.* **19**:5 (2013), 712–718. MR Zbl

[Frunzo et al. 2019] L. Frunzo, R. Garra, A. Giusti, and V. Luongo, "Modeling biological systems with an improved fractional Gompertz law", *Commun. Nonlinear Sci. Numer. Simul.* **74** (2019), 260–267. MR Zbl

[Gompertz 1825] B. Gompertz, "On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies", *Phil. Trans. R. Soc. Lond.* **115** (1825), 513–583.

[Goodrich and Peterson 2015] C. Goodrich and A. C. Peterson, *Discrete fractional calculus*, Springer, 2015. MR Zbl

[Hilling et al. 2016a] C. Hilling, S. Welsh, and D. Smith, "Data from: age, growth and fall diet of channel catfish in Cheat Lake, West Virginia", Dryad, dataset, 2016, available at https://doi.org/10.5061/dryad.8583t.

[Hilling et al. 2016b] C. D. Hilling, S. A. Welsh, and D. M. Smith, "Age, growth and fall diet of channel catfish in Cheat Lake, West Virginia", *J. Fish. Wildl. Manag.* **7**:2 (2016), 304–314.

[Jane et al. 2020] S. A. Jane, F. A. Fernandes, E. M. Silva, J. A. Muniz, T. J. Fernandes, and G. V. Pimentel, "Adjusting the growth curve of sugarcane varieties using nonlinear models", *Ciência Rural* **50**:3 (2020), art. id. e20190408.

[Kelley and Peterson 2010] W. G. Kelley and A. C. Peterson, *The theory of differential equations: classical and qualitative*, 2nd ed., Springer, 2010. MR Zbl

[Kuttner 1957] B. Kuttner, "On differences of fractional order", *Proc. London Math. Soc.* (3) **7** (1957), 453–466. MR Zbl

[Laird 1964] A. K. Laird, "Dynamics of tumour growth", *British J. Cancer* **18**:3 (1964), 490–502.

[Miller and Ross 1989] K. S. Miller and B. Ross, "Fractional difference calculus", pp. 139–152 in *Univalent functions, fractional calculus, and their applications* (Kōriyama, 1988), edited by H. M. Srivastava and S. Owa, Horwood, Chichester, 1989. MR Zbl

[Nobile et al. 1982] A. G. Nobile, L. M. Ricciardi, and L. Sacerdote, "On Gompertz growth model and related difference equations", *Biol. Cybernetics* **42** (1982), 221–229. Zbl

[Pauly and Morgan 1987] D. Pauly and G. R. Morgan (editors), *Length-based methods in fisheries research* (Sicily, Italy, 1985), ICLARM Conference Proceedings **13**, International Center for Living Aquatic Resources Management, Safat, Kuwait, 1987.

[Pezzini et al. 2019] R. V. Pezzini, A. C. Filho, F. Carini, C. T. Bandeira, J. A. Kleinpaul, and D. L. Silveira, "Modeling the growth of sudangrass cultivars at sowing times", *J. Agric. Sci.* **11**:14 (2019), 84–95.

[Ross 1975] B. Ross, "A brief history and exposition of the fundamental theory of fractional calculus", pp. 1–36 in *Fractional calculus and its applications*, edited by B. Ross, Lecture Notes in Mathematics **457**, Springer, 1975. Zbl

[Satoh 2000] D. Satoh, "A discrete Gompertz equation and a software reliability growth model", *IEICE Trans. Info. Sys.* **E83-D**:7 (2000), 1508–1513.

[Solís-Pérez et al. 2019] J. E. Solís-Pérez, J. F. Gómez-Aguilar, R. F. Escobar-Jiménez, L. Torres, and V. H. Olivares-Peregrino, "Parameter estimation of fractional Gompertz model using cuckoo search algorithm", pp. 81–95 in *Fractional derivatives with Mittag–Leffler kernel*, edited by J. F. Gómez et al., Stud. Syst. Decis. Control **194**, Springer, 2019. MR Zbl

[Wang et al. 2014] J. Wang, H. Xiang, and F. Chen, "Existence of positive solutions for a discrete fractional boundary value problem", *Adv. Difference Equ.* (2014), art. id. 253. MR Zbl

[Winsor 1932] C. P. Winsor, "The Gompertz curve as a growth curve", *Proc. Natl. Acad. Sci. USA* **18**:1 (1932), 1–8. Zbl

[Wu et al. 2015] G.-C. Wu, D. Baleanu, S.-D. Zeng, and Z.-G. Deng, "Discrete fractional diffusion equation", *Nonlinear Dynam.* **80**:1-2 (2015), 281–286. MR Zbl

[Yang et al. 2019a] C. Yang, J. Zhao, R. E. Diaz, and N. Lyu, "Data from: development of sexual dimorphism in two sympatric skinks with different growth rates", Dryad, dataset, 2019, available at https://doi.org/10.5061/dryad.nb25502.

[Yang et al. 2019b] C. Yang, J. Zhao, R. E. Diaz, and N. Lyu, "Development of sexual dimorphism in two sympatric skinks with different growth rates", *Ecol. Evol.* **9**:13 (2019), 7752–7760.

Received: 2020-04-29 Revised: 2020-08-06 Accepted: 2020-08-10

tcuchta@fairmontstate.edu *Department of Computer Science and Mathematics, Fairmont State University, Fairmont, WV, United States*

brookefincham@gmail.com *Department of Computer Science and Mathematics, Fairmont State University, Fairmont, WV, United States*

Guidelines for Authors

Submissions in all mathematical areas are encouraged. All manuscripts accepted for publication in *Involve* are considered publishable in quality journals in their respective fields, and include a minimum of one-third student authorship. Submissions should include substantial faculty input; faculty co-authorship is strongly encouraged. Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

involve
2020 vol. 13 no. 4