involve

a journal of mathematics

mathematical sciences publishers

2011 vol. 4, no. 4

involve

msp.berkeley.edu/involve

PRODUCTION

Silvio Levy, Scientific Editor   Sheila Newbery, Senior Production Editor   Cover design: ©2008 Alex Scorpan

See inside back cover or http://msp.berkeley.edu/involve for submission instructions.

The subscription price for 2011 is US $100/year for the electronic version, and $130/year (+$35 shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94704-3840, USA.

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

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

PUBLISHED BY
mathematical sciences publishers
http://msp.org/
A NON-PROFIT CORPORATION
Typeset in LaTeX

INVOLVE 4:4(2011)

Maximality of the Bernstein polynomials

Christopher Frayer and Christopher Shafhauser

(Communicated by Martin Bohner)

For fixed a and b, let Q_n be the family of polynomials $q(x)$ all of whose roots are real numbers in $[a, b]$ (possibly repeated), and such that $q(a) = q(b) = 0$. Since an element of Q_n is completely determined by it roots (with multiplicity), we may ask how the polynomial is sensitive to changes in the location of its roots. It has been shown that one of the Bernstein polynomials $b_i(x) = (x-a)^{n-i}(x-b)^i$, $i = 1, \dots, n-1$, is the member of Q_n with largest supremum norm in $[a, b]$. Here we show that for $p \geq 1$, $b_1(x)$ and $b_{n-1}(x)$ are the members of Q_n that maximize the L^p norm in $[a, b]$. We then find the associated maximum values.

1. Introduction

A monic polynomial $q(x)$ is completely determined by its roots (with multiplicity), since it can be written as the product

$$q(x) = \prod_{i=1}^{n}(x - r_i),$$

where the r_i are the roots. So it is a fair question to ask how the polynomial q is sensitive to changes in the location of its roots. Boelkins, Miller and Vugteveen [Boelkins et al. 2006] have shown that, among degree-n monic polynomials $q(x)$ all of whose roots are real, belong to $[a, b]$, and include a and b, the value of the supremum norm, $\max_{a \leq x \leq b} q(x)$, is maximized by the polynomials

$$(x-a)^{n-1}(x-b) \quad \text{and} \quad (x-a)(x-b)^{n-1}.$$

So these are in some sense the “largest” polynomials in the class just described.

We will show that these are also the largest polynomials with respect to another measure of size, namely, the L^p norm for $p \geq 1$. (For $p = 1$ this is simply the area enclosed by the graph between a and b.)

MSC2000: 30C15.

Keywords: polynomial root dragging, L^p norm, Bernstein polynomial.

Throughout this paper we let $q(x)$ be a monic polynomial of degree n all of whose roots are real and lie in $[a, b]$; we assume further that $q(a) = q(b) = 0$. We denote the family of all such polynomials by Q_n. We show that given any $q \in Q_n$,

$$\int_a^b |q(x)|\ dx \leq (b-a)^{n+1} \frac{1}{n(n+1)},$$

and for any $p \in \mathbb{N}$

$$\int_a^b |q(x)|^p\ dx \leq (b-a)^{pn+1} \frac{1}{pn+1} \left(\frac{(pn-p)!\,p!}{(pn)!} \right).$$

We then use these bounds to verify the results of [Boelkins et al. 2006]. That is, for $a < x < b$,

$$|q(x)| \leq \frac{(b-a)^n}{n} \left(\frac{n-1}{n} \right)^{n-1}.$$

2. Preliminary information

We are interested in how "large" a polynomial in Q_n can be and therefore need a way to tell when one polynomial is larger than another. We will use the L^p norms to measure the size of a polynomial. Given a polynomial q we use the notation $\|q\|_{L^p_{[a,b]}}$ to denote the L^p norm of q:

$$\|q\|_{L^p_{[a,b]}} = \left(\int_a^b |q(x)|^p \, dx \right)^{1/p}$$

and

$$\|q\|_{L^\infty_{[a,b]}} = \max_{x \in [a,b]} |q(x)|.$$

In particular, the L^1 norm of q,

$$\|q\|_{L^1_{[a,b]}} = \int_a^b |q(x)| \, dx,$$

measures the area enclosed by q.

Our goal is to understand how the L^p norm of $q \in Q_n$ is a function of the location of its roots. Specifically, we would like to understand how the smallest root of q which is greater than a will affect the L^p norm of q. We let $r_0 = a$ and r_1 represent the smallest root greater than r_0. With this in mind, we study how r_1 affects the L^p norm of polynomials of the form

$$q(x) = (x - r_1)^k s(x)$$

where $s(x) = (x - r_0)^l (x - r_2)(x - r_3) \cdots (x - r_{m-1})$ and $n = l + k + m - 2$. That is, q is a degree n polynomial with roots

$$r_0 = a < r_1 < r_2 \leq r_3 \cdots \leq r_{m-1} = b,$$

which takes into account having possibly repeated roots at r_0 and r_1. To understand how r_1 affects the L^p norm of q we study the function

$$A_p(q)(r_1) = \|q\|^p_{L^p_{[a,b]}} = \int_a^{r_1} (r_1 - x)^{kp} |s(x)|^p \, dx + \int_{r_1}^b (x - r_1)^{kp} |s(x)|^p \, dx,$$

where we allow $r_1 \in [r_0, r_2]$.

The following two basic results of calculus will be used later, when we optimize the L^p norm.

Lemma 2.1. *If $f(x)$ is twice differentiable and concave up on $[a, b]$, then*

$$\max\{f(a), f(b)\} > f(x)$$

for all $x \in (a, b)$.

Lemma 2.2 (Leibniz's formula). *If $F(x, y)$ and $F_x(x, y)$ are continuous in both x and y in some region of the xy-plane including $a \leq y \leq x$ and $u(x)$ is a continuous function of x, then*

$$\frac{d}{dx} \int_a^{u(x)} F(x, y) \, dy = F(x, u(x)) \frac{d}{dx} u(x) + \int_a^{u(x)} F_x(x, y) \, dy.$$

3. Maximizing the enclosed area

We are now ready to find the member of Q_n that encloses the largest area. In order to do so we show that $A_1(q)(r_1)$ is concave up on $[r_0, r_2]$.

Theorem 3.1. *If $q(x) = (x - r_1)^k s(x)$, where $s(x) = (x - r_0)^l (x - r_2) \cdots (x - r_{m-1})$ and $r_0 < r_1 < r_2 \leq r_3 \leq \cdots \leq r_{m-1}$, then*

$$\frac{d^2}{dr_1^2} A_1(q)(r_1) > 0 \quad \textit{on } [r_0, r_2].$$

Proof. Let $F(r_1, x) = (x - r_1)^k s(x)$, and observe that $F(r_1, r_1) = 0$. Applying Leibniz's formula to each term in $dA_1(q)(r_1)/dr_1$, we have

$$\frac{d}{dr_1} \int_a^{r_1} (r_1 - x)^k |s(x)| \, dx = k \int_a^{r_1} (r_1 - x)^{k-1} |s(x)| \, dx$$

and

$$\frac{d}{dr_1} \int_{r_1}^b (x - r_1)^k |s(x)| \, dx = -k \int_{r_1}^b (x - r_1)^{k-1} |s(x)| \, dx.$$

If $k = 1$, the fundamental theorem of Calculus implies that

$$\frac{d^2}{dr_1^2}\int_a^{r_1}(r_1 - x)|s(x)|\,dx = |s(r_1)| \quad \text{and} \quad \frac{d^2}{dr_1^2}\int_{r_1}^{b}(x - r_1)|s(x)|\,dx = |s(r_1)|.$$

Since r_1 is not a root of $s(x)$, it follows that

$$\frac{d^2}{dr_1^2}A_1(q)(r_1) = 2|s(r_1)| > 0.$$

If $k \geq 2$, then

$$\frac{d^2}{dr_1^2}\int_a^{r_1}(r_1 - x)^k|s(x)|\,dx = k(k-1)\int_a^{r_1}(r_1 - x)^{k-2}|s(x)|\,dx$$

and

$$\frac{d^2}{dr_1^2}\int_{r_1}^{b}(x - r_1)^k|s(x)|\,dx = k(k-1)\int_{r_1}^{b}(x - r_1)^{k-2}|s(x)|\,dx.$$

Therefore,

$$\frac{d^2}{dr_1^2}A_1(q)(r_1) = k(k-1)\int_a^{b}|(x - r_1)^{k-2}s(x)|\,dx > 0$$

and $A_1(q)(r_1)$ is concave up on $[r_0, r_2]$. □

Corollary 3.2. *One of the Bernstein polynomials*

$$b_i(x) = (x-a)^{n-i}(x-b)^i, \quad i = 1, \dots, n-1,$$

is the member of Q_n that encloses the largest area on $[a, b]$.

Theorem 3.1, along with Lemma 2.1, tells us that we can always find a polynomial in Q_n with a larger L^1 norm by "dragging" r_1 to either r_0 or r_2. Playing this game a finite number of times leaves us a polynomial with roots only at a and b. So, one of the Bernstein polynomials,

$$b_i(x) = (x-a)^{n-i}(x-b)^i, \quad i = 1, \dots, n-1,$$

will be the member of Q_n that encloses the largest area.

4. Other values of *p*

We now extend the method of the previous section to values of $p > 1$. Let

$$q(x) = (x - r_1)^k s(x),$$

where

$$s(x) = (x - r_0)^l (x - r_2) \cdots (x - r_{m-1})$$

with $r_0 < r_1 < r_2 \leq r_3 \leq \cdots \leq r_{m-1}$, and consider

$$A_p(q)(r_1) = \int_a^{r_1} (r_1 - x)^{kp} |s(x)|^p \, dx + \int_{r_1}^b (x - r_1)^{kp} |s(x)|^p \, dx. \tag{1}$$

If we can show that $A_p(q)(r_1)$ is concave up on $[r_0, r_2]$, then one of the Bernstein polynomials will be the member of Q_n with the largest L^p norm. Using the same argument as the $p = 1$ case, two applications of Leibniz's formula yields

$$\frac{d^2}{dr_1^2} A_p(q)(r_1) = kp(kp-1) \int_a^b |(x - r_1)^{kp-2}||s(x)|^p \, dx > 0,$$

and $A_p(q)(r_1)$ is concave up on the interval $[r_0, r_2]$ when $p > 1$.

In the above calculation, we have to be careful when $kp-2 < 0$. Since $kp-1 > 0$ ($k \geq 1$ and $p > 1$) the hypothesis of Leibniz's formula are satisfied for the first application with

$$\frac{d}{dr_1} A_p(q)(r_1) = kp \int_a^{r_1} (r_1 - x)^{kp-1} |s(x)|^p \, dx - kp \int_{r_1}^b (x - r_1)^{kp-1} |s(x)|^p \, dx. \tag{2}$$

When applying Leibniz's formula to the first term on the right-hand side, we need

$$\frac{\partial}{\partial r_1} (r_1 - x)^{kp-1} |s(x)|^p$$

to be continuous in both x and r_1 in some region including $a \leq x \leq r_1$. Although this may not be true at $x = r_1$, we can still justify the application of Leibniz's formula by considering the interval $[a, r_1 - \epsilon]$ and letting $\epsilon \to 0^+$. That is,

$$\frac{d^2}{dr_1^2} \int_a^{r_1} (r_1 - x)^{kp} |s(x)|^p \, dx = \lim_{\epsilon \to 0^+} \left(\frac{d}{dr_1} kp \int_a^{r_1 - \epsilon} (r_1 - x)^{kp-1} |s(x)|^p \, dx \right).$$

Because the integrand is positive, the result will follow if the limit exists.

The polynomial $s(x)$ does not change sign on the interval (a, r_2), so we may assume without loss of generality that $s(x) \geq 0$ on $[a, r_1 - \epsilon]$, with $s(x) = 0$ only at $x = a$. Applying Leibniz's formula on $[a, r_1 - \epsilon]$ yields

$$\begin{aligned}
&\lim_{\epsilon \to 0^+} \left(\frac{d}{dr_1} kp \int_a^{r_1 - \epsilon} (r_1 - x)^{kp-1} s(x)^p \, dx \right) \\
&\quad = \lim_{\epsilon \to 0^+} kp(kp-1) \int_a^{r_1 - \epsilon} (r_1 - x)^{kp-2} s(x)^p \, dx + \lim_{\epsilon \to 0^+} (\epsilon)^{kp-1} s(r_1 - \epsilon)^p \\
&\quad = \lim_{\epsilon \to 0^+} kp(kp-1) \int_a^{r_1 - \epsilon} (r_1 - x)^{kp-2} s(x)^p \, dx.
\end{aligned}$$

In order to see that this limit exists, we integrate by parts to get

$$\begin{aligned} kp(kp-1)\lim_{\epsilon\to 0^+}\left(-s(r_1-\epsilon)^p\frac{(\epsilon)^{kp-1}}{kp-1}+\frac{p}{kp-1}\int_a^{r_1-\epsilon}(r_1-x)^{kp-1}s(x)^{p-1}s'(x)\,dx\right) \\ = kp^2\int_a^{r_1}(r_1-x)^{kp-1}s(x)^{p-1}s'(x)\,dx, \end{aligned}$$

where equality follows as $kp-1>0$ and the integrand is a continuous function of x on $[a, r_1]$. Hence the limit exists and is positive from an earlier observation. A similar argument applied to the second term on the right in (2) shows that

$$\frac{d^2}{dr_1^2}\int_{r_1}^{b}(x-r_1)^{kp}|s(x)|^p\,dx = \lim_{\epsilon\to 0^+}\frac{d}{dr_1}\left(-kp\int_{r_1+\epsilon}^{b}(x-r_1)^{kp-1}|s(x)|^p\,dx\right)$$

exists and is positive. Therefore, $\dfrac{d^2}{dr_1^2}A_p(q)(r_1)>0$.

From an argument similar to Theorem 3.1, we have the following result:

Theorem 4.1. *If $p \geq 1$, one of the Bernstein polynomials is the member of Q_n that has the largest L^p norm on $[a, b]$.*

Finally, we consider the case $p=\infty$. Since $[a, b]$ has finite measure,

$$\lim_{p\to\infty}\|f(x)\|_{L^p_{[a,b]}} = \|f(x)\|_{L^\infty_{[a,b]}}; \tag{3}$$

see [Wheeden and Zygmund 1977, p. 126].

Corollary 4.2. *One of the Bernstein polynomials is the member of Q_n that has the largest L^∞ norm on $[a, b]$.*

Proof. Let $m(x)\in Q_n$ with $m(x)\neq b_i(x)$ for $i=1,\dots,n-1$. If we restrict p to the positive integers, it follows from (3) that the sequences

$$\left\{\|m(x)\|_{L^p_{[a,b]}}\right\}_p \to \|m(x)\|_{L^\infty_{[a,b]}} \quad\text{and}\quad \left\{\|b_i(x)\|_{L^p_{[a,b]}}\right\}_p \to \|b_i(x)\|_{L^\infty_{[a,b]}}$$

as $p\to\infty$. Theorem 4.1 implies that for each $p\in\mathbb{N}$

$$\|m(x)\|_{L^p_{[a,b]}} \leq \|b_i(x)\|_{L^p_{[a,b]}},$$

so that

$$\lim_{p\to\infty}\|m(x)\|_{L^p_{[a,b]}} \leq \lim_{p\to\infty}\|b_i(x)\|_{L^p_{[a,b]}}.$$

Therefore $\|m(x)\|_{L^\infty_{[a,b]}} \leq \|b_i(x)\|_{L^\infty_{[a,b]}}$ and we have the desired result. □

5. Evaluating the maximum

The process of increasing the L^p norm lead us to a finite class of polynomials that must contain the "largest" polynomial in Q_n. Specifically, we arrived at the class of Bernstein polynomials

$$b_i(x) = (x-a)^{n-i}(x-b)^i, \quad i = 1, \dots, n-1.$$

We would like to determine which of these polynomials will maximize the L^p norm. To do so, we recall (from [Dennery and Krzywicki 1996, pp. 94–98], for example) the beta function, defined by

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)},$$

where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t} dt$ satisfies the property $\Gamma(n+1) = n!$.

Initially, we answer the question when $a = 0$ and $b = 1$, and then translate the result back to general a and b by the appropriate substitution. We observe that

$$\int_0^1 x^{n-i}(x-1)^i \, dx = (-1)^i B(n-i+1, i+1) = (-1)^i \frac{\Gamma(n-i+1)\Gamma(i+1)}{\Gamma(n+2)}.$$

Since the polynomials $b_i(x)$ are either entirely positive or entirely negative on [0, 1], we have

$$\|b_i(x)\|_{L^1_{[0,1]}} = \left| \int_0^1 x^{n-i}(x-1)^i \, dx \right| = \frac{\Gamma(n-i+1)\Gamma(i+1)}{\Gamma(n+2)} = \frac{1}{n+1} \frac{i!\,(n-i)!}{n!}.$$

Note that $\dfrac{i!\,(n-i)!}{n!}$ is the reciprocal of the binomial coefficient $\dbinom{n}{i}$. Since n is fixed, we need to pick the value of i that minimizes this binomial coefficient. Clearly this happens when $i = 1$ or $i = n-1$. Therefore, the maximum value of the norm is obtained for $b_1(x)$ and $b_{n-1}(x)$:

$$\|b_1(x)\|_{L^1_{[0,1]}} = \|b_{n-1}(x)\|_{L^1_{[0,1]}} = \frac{1}{n+1}\binom{n}{1}^{-1} = \frac{1}{n(n+1)}. \tag{4}$$

This can be generalized to the interval $[a, b]$ by using the substitution $u = (x-a)/(b-a)$; for any monic degree-n polynomial $q(x)$ with all real zeros in $[a, b]$ such that $q(x)$ has roots at a and b, we have

$$\|q(x)\|_{L^1_{[a,b]}} \le (b-a)^{n+1} \frac{1}{n(n+1)}.$$

If $p \in \mathbb{N}$, the same method can be used to evaluate the L^p norm of the Bernstein polynomials. We have

$$\|b_i(x)\|^p_{L^p_{[0,1]}} = \left[\frac{\Gamma(pn-pi+1)\Gamma(pi+1)}{\Gamma(pn+2)}\right]^{1/p} = \left[\frac{1}{pn+1}\,\frac{(pn-pi)!\,(pi)!}{(pn)!}\right]^{1/p}. \quad (5)$$

The maximum value is still achieved by $b_1(x)$ and $b_{n-1}(x)$. Inequality (5) can be generalized to the interval $[a, b]$ by using the substitution $u = (x-a)/(b-a)$; for any monic degree-n polynomial $q(x)$ with all real zeros in $[a, b]$ such that $q(x)$ has roots at a and b,

$$\|q(x)\|_{L^p_{[a,b]}} \le \left[(b-a)^{pn+1}\frac{1}{pn+1}\,\frac{(pn-pi)!\,(pi)!}{(pn)!}\right]^{1/p}.$$

If p is not a natural number, the first equality in (5) is still valid, though we can no longer express the result in terms of factorials. Therefore (again passing to the case of $[a, b]$) we can write

$$\|b_i(x)\|_{L^p_{[a,b]}} = \left[(b-a)^{pn+1}\frac{\Gamma(pn-pi+1)\Gamma(pi+1)}{\Gamma(pn+2)}\right]^{1/p}. \quad (6)$$

To find the values of i that maximize this expression, we can differentiate it with respect to i. (Although only integer values of i make sense in our context, the quotient in (6) makes sense for all real i in the range of interest, $1 \le i \le n-1$. The domain of definition and differentiability of the gamma function includes $(0, \infty)$.) The derivative of the gamma function involves another transcendental function, known as polygamma. The upshot is that the quotient in (6) has only one critical point in the interval $1 \le i \le n-1$, and it is a minimum rather than a maximum. It follows that, once more, the local maxima in this interval must be at the endpoints of the interval, that is, $i = 1$ and $i = n-1$.

6. Recovering the supremum norm

As mentioned in the introduction, it was established in [Boelkins et al. 2006] that the Bernstein polynomials $b_1(x)$ and $b_{n-1}(x)$ are the members of Q_n with the largest L^∞ norm on $[a, b]$. In fact, they found that

$$\|b_1(x)\|_{L^\infty_{[a,b]}} = \frac{(b-a)^n}{n}\left(\frac{n-1}{n}\right)^{n-1},$$

a result that we now reproduce as a consequence of the work in the previous section.

We have seen that, for $p \in \mathbb{N}$,

$$\|b_1(x)\|_{L^p_{[a,b]}} = \left[\frac{(b-a)^{pn+1}}{pn+1}\left(\frac{(pn-p)!\,p!}{(pn)!}\right)\right]^{1/p}.$$

Applying Sterling's approximation, $\lim_{n\to\infty}\left(n! - \sqrt{2\pi n}\left(\frac{n}{e}\right)^n\right) = 0$, we obtain

$$\begin{aligned}\|b_1(x)\|_{L^\infty_{[a,b]}} &= \lim_{p\to\infty}\|b_1(x)\|_{L^p_{[a,b]}}\\ &= \lim_{p\to\infty}\left[\frac{(b-a)^{pn+1}}{pn+1}\frac{(pn-p)!\,p!}{(pn)!}\right]^{1/p}\\ &= \lim_{p\to\infty}\left[\frac{(b-a)^{pn+1}}{pn+1}\frac{\sqrt{2\pi p(n-1)}\left(\frac{p(n-1)}{e}\right)^{p(n-1)}\sqrt{2\pi p}\left(\frac{p}{e}\right)^p}{\sqrt{2\pi pn}\left(\frac{pn}{e}\right)^{pn}}\right]^{1/p}.\end{aligned}$$

After simplification, this becomes

$$\begin{aligned}\|b_1(x)\|_{L^\infty_{[a,b]}} &= \frac{(b-a)^n}{n}\left(\frac{n-1}{n}\right)^{n-1}\lim_{p\to\infty}\left[\frac{(b-a)}{pn+1}\left(\frac{\sqrt{2\pi p(n-1)}}{\sqrt{n}}\right)\right]^{1/p}\\ &= \frac{(b-a)^n}{n}\left(\frac{n-1}{n}\right)^{n-1}\lim_{p\to\infty}\left(\frac{b-a}{\sqrt{n}}\right)^{1/p}\lim_{p\to\infty}\left(\frac{\sqrt{2\pi p(n-1)}}{pn+1}\right)^{1/p}\\ &= \frac{(b-a)^n}{n}\left(\frac{n-1}{n}\right)^{n-1}\lim_{p\to\infty}\left(\frac{\sqrt{2\pi p(n-1)}}{pn+1}\right)^{1/p}.\end{aligned}$$

L'Hopital's rule implies

$$\lim_{p\to\infty}\left(\frac{\sqrt{2\pi p(n-1)}}{pn+1}\right)^{1/p} = 1$$

and it follows that

$$\|b_1(x)\|_{L^\infty_{[a,b]}} = \frac{(b-a)^n}{n}\left(\frac{n-1}{n}\right)^{n-1}.$$

We can now reasonably claim that the Bernstein polynomials are the largest monic polynomials with all real roots in $[a, b]$ in the full sense of all possible L^p norms.

References

[Boelkins et al. 2006] M. Boelkins, J. Miller, and B. Vugteveen, "From Chebyshev to Bernstein: a tour of polynomials small and large", *College Math. J.* **37**:3 (2006), 194–204. MR 2007i:12001

[Dennery and Krzywicki 1996] P. Dennery and A. Krzywicki, *Mathematics for physicists*, Dover, Mineola, NY, 1996. MR 42 #1376 Zbl 1141.00003

[Wheeden and Zygmund 1977] R. L. Wheeden and A. Zygmund, *Measure and integral: an introduction to real analysis*, Pure and Applied Mathematics **43**, Marcel Dekker, New York, 1977. MR 58 #11295 Zbl 0362.26004

Received: 2010-05-05 Revised: 2011-05-04 Accepted: 2011-07-12

frayerc@uwplatt.edu *Mathematics Department, University of Wisconsin-Platteville, 1 University Plaza, Platteville, WI 53818, United States*

shafhauserc@uwplatt.edu *Department of Mathematics, University of Nebraska-Lincoln, Lincoln, NE 68588, United States*

INVOLVE 4:4(2011)

The family of ternary cyclotomic polynomials with one free prime

Yves Gallot, Pieter Moree and Robert Wilms

(Communicated by Kenneth S. Berenhaut)

A cyclotomic polynomial $\Phi_n(x)$ is said to be ternary if $n = pqr$, with p, q and r distinct odd primes. Ternary cyclotomic polynomials are the simplest ones for which the behavior of the coefficients is not completely understood. Here we establish some results and formulate some conjectures regarding the coefficients appearing in the polynomial family $\Phi_{pqr}(x)$ with $p < q < r$, p and q fixed and r a free prime.

1. Introduction

The n-th cyclotomic polynomial $\Phi_n(x)$ is defined by

$$\Phi_n(x) = \prod_{\substack{1\le j\le n\\(j,n)=1}} (x-\zeta_n^j) = \sum_{k=0}^{\infty} a_n(k)x^k,$$

with ζ_n a n-th primitive root of unity (one can take $\zeta_n = e^{2\pi i/n}$). It has degree $\varphi(n)$, with φ Euler's totient function. We write $A(n) = \max\{|a_n(k)| : k \ge 0\}$, and this quantity is called the height of $\Phi_n(x)$. It is easy to see that $A(n) = A(N)$, with $N = \prod_{p|n,\ p>2} p$ the odd squarefree kernel. In deriving this, one uses the observation that if n is odd, then $A(2n) = A(n)$. If n has at most two distinct odd prime factors, then $A(n) = 1$. If $A(n) > 1$, then we necessarily must have that n has at least three distinct odd prime factors. In particular for $n < 105 = 3\cdot 5\cdot 7$ we have $A(n) = 1$. It turns out that $A(105) = 2$ with $a_{105}(7) = -2$. Thus the easiest case where we can expect nontrivial behavior of the coefficients of $\Phi_n(x)$ is the ternary case, where $n = pqr$, with $2 < p < q < r$ odd primes. In this paper we are concerned with the family of ternary cyclotomic polynomials

$$\{\Phi_{pqr}(x) : r > q\}, \tag{1}$$

MSC2000: primary 11C08; secondary 11B83.

Keywords: ternary cyclotomic polynomial, coefficient.

where $2 < p < q$ are fixed primes and r is a "free prime". Up to now in the literature the above family was considered, but with also q free. The maximum coefficient (in absolute value) that occurs in that family will be denoted by $M(p)$, thus $M(p) = \max\{A(pqr) : p < q < r\}$, with $p > 2$ fixed. Similarly we define $M(p; q)$ to be the maximum coefficient (in absolute value) that occurs in the family (1), thus $M(p; q) = \max\{A(pqr) : r > q\}$, with $2 < p < q$ fixed primes.

Example. Bang [1895] proved that $M(p) \le p - 1$. Since $a_{3\cdot 5\cdot 7}(7) = -2$ we infer that $M(3) = 2$. Using $a_{105}(7) = -2$ and $M(3) = 2$, we infer that $M(3; 5) = 2$.

Let $\mathcal{A}(p; q) = \{a_{pqr}(k) : r > q,\ k \ge 0\}$ be the set of coefficients occurring in the polynomial family (1).

Proposition 1. $\mathcal{A}(p; q) = [-M(p; q), M(p; q)] \cap \mathbb{Z}$.

This shows the relevance of understanding $M(p; q)$. Let us first recall some known results concerning the related function $M(p)$. Here we know thanks to Bachman [2003], who very slightly improved on an earlier result in [Beiter 1971], that $M(p) \le 3p/4$. It was conjectured by Sister Marion Beiter [1968] (see also [Beiter 1971]) that $M(p) \le (p+1)/2$. She proved it for $p \le 5$. Since Möller [1971] proved that $M(p) \ge (p+1)/2$ for $p > 2$, her conjecture actually would imply that $M(p) = (p+1)/2$ for $p > 2$. The first to show that Beiter's conjecture is false seems to have been Eli Leher (in his PhD thesis), who gave the counterexample $a_{17\cdot 29\cdot 41}(4801) = -10$, showing that $M(17) \ge 10 > 9 = (17+1)/2$. Gallot and Moree [2009b] provided for each $p \ge 11$ infinitely many infinitely many counterexamples $p \cdot q_j \cdot r_j$ with q_j strictly increasing with j. Moreover, they have shown that for every $\epsilon > 0$ and p sufficiently large $M(p) > (\frac{2}{3} - \epsilon)p$. They also proposed the corrected Beiter conjecture: $M(p) \le 2p/3$. The implications of their work for $M(p; q)$ are described in Section 4.

Proposition 1 together with Möller's result quoted above gives a different proof of the result, due to Bachman [2004], that $\{a_{pqr}(k) : p < q < r\} = \mathbb{Z}$. For references and further results in this direction (begun by I. Schur) see Fintzen [2011].

Jia Zhao and Xianke Zhang [2010] showed that $M(7) = 4$, thus establishing the Beiter conjecture for $p = 7$. In a later paper they established the corrected Beiter conjecture:

Theorem 2 [Zhao and Zhang 2009]. $M(p) \le 2p/3$.

This result together with some computer computation allows one to extend the list of exactly known values of $M(p)$ (see Table 1).

It is not known whether there is a finite procedure to determine $M(p)$. On the other hand, it is not difficult to see that there is such a procedure for $M(p; q)$.

Proposition 3. *Given primes $2 < p < q$, there is a finite procedure to determine $M(p; q)$.*

p	$M(p)$	smallest n
3	2	$3\cdot 5\cdot 7$
5	3	$5\cdot 7\cdot 11$
7	4	$7\cdot 17\cdot 23$
11	7	$11\cdot 19\cdot 601$
13	8	$13\cdot 73\cdot 307$
19	12	$19\cdot 53\cdot 859$

Table 1. Values of $M(p)$. By "smallest n" we mean the smallest integer n satisfying $A(n) = M(p)$ and with p as its smallest prime divisor.

Recall that a set S of primes is said to have *natural density* δ if

$$\lim_{x\to\infty} \frac{|\{p \le x : p \in S\}|}{\pi(x)} = \delta,$$

where $\pi(x)$ is the number of primes $p \le x$. A further question that arises is how often the maximum value $M(p)$ is assumed. We have:

Theorem 4. *Given primes $2 < p < q$, there exists a prime q_0 with $q_0 \equiv q \pmod p$ and an integer d such that $M(p, q) \le M(p, q_0) = M(p, q')$ for every prime $q' \ge q_0$ satisfying $q' \equiv q_0 \pmod{d\cdot p}$. In particular the set of primes q with $M(p; q) = M(p)$ has a subset having a positive natural density.*

A weaker result in this direction, namely that for a fixed prime $p \ge 11$, the set of primes q such that $M(p; q) > (p + 1)/2$ has a subset of positive natural density, follows from [Gallot and Moree 2009b] (recall that $M(p) > (p+1)/2$ for $p \ge 11$).

Unfortunately, the proof of Theorem 4 gives a lower bound for the density that seems to be far removed from the true value. In this paper we present some constructions that allow one to obtain much better bounds for the density for small p. These results are subsumed in the following main result of the paper.

Theorem 5. *Let $2 < p \le 19$ be a prime with $p \ne 17$. Then the set of primes q such that $M(p; q) = M(p)$ has a subset having natural density $\delta(p)$ as follows*:

$p =$	3	5	7	11	13	19
$\delta(p) =$	1	1	1	$\frac{2}{5}$	$\frac{1}{12}$	$\frac{1}{9}$

Numerical experimentation suggests that the set of primes q such that $M(p; q) = M(p)$ has a natural density $\delta(p)$ as given in the above table, except when $p = 13$ in which case numerical experimentation suggests $\delta(13) = 1/3$.

In order to prove Theorem 5, we will use the following theorem dealing with $2 < p \le 7$.

Theorem 6. *For $2 < p \le 7$ and $q > p$ we have $M(p;\, q) = (p+1)/2$, except in the case $p = 7$, $q = 13$, where $M(7;\, 13) = 3$.*

The fact that $M(7;\, 13) = 3$ can be explained. It turns out that if $ap + bq = 1$ for integers a and b small in absolute value, then $M(p;\, q)$ is small. For example:

Theorem 7. *If $p \ge 5$ and $2p - 1$ is a prime, then $M(p;\, 2p-1) = 3$.*

This result and similar ones are established in Section 10.

Our main conjecture on $M(p;\, q)$ is the following one.

Conjecture 8. *Given a prime p, there exists an integer d and a function*

$$g : (\mathbb{Z}/d\mathbb{Z})^* \to \mathbb{Z}_{>0}$$

such that for some $q_0 > d$ we have for every prime $q \ge q_0$ that $M(p;\, q) = g(\bar{q})$, where $1 \le \bar{q} < d$ satisfies $q \equiv \bar{q} \pmod{d}$. The function g is symmetric, that is we have $g(\alpha) = g(d - \alpha)$.

The smallest integer d with the above properties, if it exists, we call the *ternary conductor* $\mathfrak{f}_p$. The corresponding smallest choice of q_0 (obtained on setting $d = \mathfrak{f}_p$) we call the *ternary minimal prime*. For $p = 7$ we obtain, e.g., $\mathfrak{f}_7 = 1$ and $q_0 = 17$ (by Theorem 6). Note that once we know q_0 it is a finite computation to determine d and the function g. Theorem 6 can be used to obtain the $p \le 7$ part of the following observation concerning the ternary conductor.

Proposition 9. *If $2 < p \le 7$, then the ternary conductor exists and we have $\mathfrak{f}_p = 1$. If $p \ge 11$ and $\mathfrak{f}_p$ exists, then $p | \mathfrak{f}_p$.*

While Theorem 4 only says that the set of primes q with $M(p;\, q) = M(p)$ has a subset having a positive natural density, Conjecture 8 implies that the set actually has a natural density in $\mathbb{Q}_{>0}$ which can be easily explicitly computed assuming we know q_0. In order to establish this implication one can invoke a quantitative form of Dirichlet's prime number theorem to the effect that, for $(a, d) = 1$, we have, as x tends to infinity,

$$\sum_{\substack{p \le x \\ p \equiv a \pmod{d}}} 1 \sim \frac{x}{\varphi(d) \log x}. \tag{2}$$

This result implies that asymptotically the primes are equidistributed over the primitive congruence classes modulo d. (Recall that Dirichlet's prime number theorem, Dirichlet's theorem for short, says that each primitive residue class contains infinitely many primes.)

The main tool in this paper is Kaplan's lemma, presented in Section 6. The material in that section (except for Lemma 22, which is new) is taken from [Gallot and Moree 2009a]. As a demonstration of working with Kaplan's lemma two examples (with and without table) are given in Section 6.1. In [Gallot et al. 2010], the full version of this paper, details of further proofs using Kaplan's lemma can be found. In the shorter version we have merely written "Apply Kaplan's lemma".

The above summary of results makes clear how limited presently our knowledge of $M(p;q)$ is. For the benefit of the interested reader we present a list of open problems in Section 11.

2. Proof of two propositions and Theorem 4

Proof of Proposition 1. By the definition of $M(p;q)$ we have

$$\mathcal{A}(p;q) \subseteq [-M(p;q), M(p;q)] \cap \mathbb{Z}.$$

Let $r > q$ be a prime such that $A(pqr) = M(p;q)$ and suppose, without loss of generality, that $a_{pqr}(k) = M(p;q)$. Gallot and Moree [2009a] showed that $|a_n(k) - a_n(k-1)| \leq 1$ for ternary n (see [Bachman 2010; Bzdęga 2010] for alternative proofs). Since $a_{pqr}(k) = 0$ for every k large enough, it then follows that $0, 1, \ldots, M(p;q)$ are in $\mathcal{A}(p;q)$. By a result of Kaplan [2007] (see [Zhao and Zhang 2010] for a different proof), we can find a prime $s \equiv -r \pmod{pq}$ and an integer k_1 such that $a_{pqs}(k_1) = -M(p;q)$. By a similar arguments as above one then infers that $-M(p;q), -M(p;q)+1, \ldots, -1, 0$ are all in $\mathcal{A}(p;q)$. □

Proof of Proposition 3. Let $\mathcal{R}_{pq}$ be a set of primes, all exceeding q such that every primitive residue class modulo pq is represented. By [Kaplan 2007, Theorem 2] we have $A(pqr) = A(pqs)$ if $s \equiv r \pmod{pq}$ with s, r both primes exceeding q and hence

$$M(p;q) = \max\{A(pqr) : r \in \mathcal{R}_{pq}\}.$$

Since the computation of $\mathcal{R}_{pq}$ and $A(pqr)$ is a finite one, the computation of $M(p;q)$ is also finite. □

The remainder of the section is devoted to the proof of Theorem 4.

For coprime positive (not necessary prime) integers p, q, r we define

$$\Phi'_{p,q,r}(x) = \frac{(x^{pqr}-1)(x^p-1)(x^q-1)(x^r-1)}{(x-1)(x^{pq}-1)(x^{pr}-1)(x^{qr}-1)} = \sum_{k=0}^{\infty} a'_{p,q,r}(k)x^k.$$

Here we do not assume $p < q < r$. Hence we have the symmetry $\Phi'_{p,q,r}(x) = \Phi'_{p,r,q}(x)$. A routine application of the inclusion-exclusion principle to the roots of the factors shows that $\Phi'_{p,q,r}(x)$ is a polynomial. It is referred to as a ternary inclusion-exclusion polynomial. Inclusion-exclusion polynomials can be defined in great generality, and the reader is referred to [Bachman 2010] for an introductory discussion. He shows that such polynomials and thus $\Phi'_{p,q,r}(x)$ in particular, can be written as products of cyclotomic polynomials (see Theorem 2 in that reference).

Analogously to $A(pqr)$ and $M(p;q)$ we define

$$A'(p,q,r)=\max\{|a'_{p,q,r}(k)|:k\geq 0\},$$
$$M'(p;q)=\max\{A'(p,q,r):r\geq 1\},$$
$$M'(p)=\max\{M'(p;q):q\geq 1\}.$$

We have $\Phi_{pqr}(x)=\Phi'_{p,q,r}(x)$ if p,q,r are distinct primes, so $A(pqr)=A'(p,q,r)$ in this case.

Lemma 10. *For coprime positive* (*not necessary prime*) *integers p,q,r we have $A'(p,q,r_1)\leq A'(p,q,r_2)\leq A'(p,q,r_1)+1$ if $r_2\equiv r_1$* (mod pq) *and $r_2>r_1$.*

Proof. Note that $r_2>\max\{p,q\}$. If $r_1>\max\{p,q\}$, then Kaplan [2007, proof of Theorem 2] showed that $A'(p,q,r_1)=A'(p,q,r_2)$. In the remaining case $r_1<\max\{p,q\}$, we have $A'(p,q,r_1)\leq A'(p,q,r_2)\leq A'(p,q,r_1)+1$ by the Theorem in [Bachman and Moree 2011]. □

In [Bachman and Moree 2011] it is remarked that $A'(p,q,r_2)=A'(p,q,r_1)+1$ can occur.

Lemma 11. *If p is a prime, then $M'(p)=M(p)$. If q is also a prime with $q>p$ then $M'(p;q)=M(p;q)$.*

Proof. Let $p<q$ be primes. Assume $M'(p;q)=A'(p,q,r)$, where r is not necessary a prime. By Dirichlet's theorem we can find a prime r' satisfying

$$r'\equiv r(\text{mod } pq) \quad \text{and} \quad r'>\max(q,r).$$

Therefore we have, by Lemma 10,

$$M'(p;q)=A'(p,q,r)\leq A'(p,q,r')=A(p,q,r')\leq M(p;q).$$

Since obviously $M(p;q)\leq M'(p;q)$, we have $M'(p;q)=M(p;q)$.

Now let only p be a prime. Assume $M'(p)=A'(p,q,r)$, where q and r are not necessary primes. Again by Dirichlet's theorem we find a prime q' with $q'\equiv q$ (mod pr) and $q'>\max(p,q)$. Using Lemma 10 we have

$$M'(p)=A'(p,q,r)\leq A'(p,q',r)\leq M'(p,q')=M(p,q')\leq M(p).$$

Since obviously $M(p)\leq M'(p)$, we have $M'(p)=M(p)$. □

Proof of Theorem 4. We set $q_1:=q$. Let r_i be a positive integer satisfying $M'(p;q_i)=A'(p,q_i,r_i)$. Using Lemma 10 (note that $A'(p,q,r)$ is invariant under permutations of p, q and r) we deduce

$$M'(p;q_1)=A'(p,q_1,r_1)\leq A'(p,q_2,r_1)\leq A'(p,q_2,r_2)=M'(p,q_2),$$

where $q_2 = q_1 + pr_1$. By the same argument the sequence $q_1, q_2, q_3, \dots$ with $q_{i+1} = q_i + pr_i$ satisfies

$$M'(p; q_1) \le M'(p; q_2) \le M'(p; q_3) \le \cdots$$

Since $M'(p; q) \le M'(p) = M(p)$ and by, e.g., Lemma 18, $M(p)$ is finite, there are only finitely many different values for $M'(p; q)$. Hence there is an index k such that $M'(p; q_k) = M'(p; q_{k+i})$ for all $i \ge 0$. That means

$$M'(p; q_k) = A'(p, q_k, r_k) = A'(p, q_{k+1}, r_k) = A'(p, q_{k+1}, r_{k+1}) = M'(p, q_{k+1}),$$

and by induction $A'(p, q_{k+i}, r_k) = A'(p, q_{k+i}, r_{k+i})$. Therefore we can assume $r_{k+i} = r_k$ for $i \ge 0$. Then we have $q_{k+i} = q_k + i \cdot pr_k$. We set $q_0 := q_k$ and $d := r_k$. Certainly we have $q_0 \equiv q \pmod{p}$. Let $q' \ge q_0$ be a prime with $q' \equiv q_0 \pmod{d \cdot p}$. There must be an integer m such that $q' = q_{k+m}$. Since $M'(p; q) = M(p; q)$ by Lemma 11, we have

$$M(p; q_1) \le M(p; q_0) = M(p; q').$$

Applying this to $M(p; q_1)$ with $M(p; q_1) = M(p)$, where we have chosen q_1 such that $M(p; q_1) = M(p)$, we get infinitely many primes of the form $q_i = q_1 + i \cdot pr_1$ satisfying $M(p; q_i) = M(p)$. On invoking (2) with $a = q_1$ and $d = pr_1$ the proof is then completed. □

3. The bounds of Bachman and Bzdęga

Let q^* and r^*, $0 < q^*, r^* < p$ be the inverses of q and r modulo p respectively. Set $a = \min(q^*, r^*, p - q^*, p - r^*)$. Put $b = \max(\min(q^*, p - q^*), \min(r^*, p - r^*))$. In the sequel we will use repeatedly that $b \ge a$. Bachman [2003] showed that

$$A(pqr) \le \min\left(\frac{p-1}{2} + a,\ p - b\right). \tag{3}$$

This was more recently improved by Bzdęga [Bzdęga 2010] who showed that

$$A(pqr) \le \min(2a + b,\ p - b). \tag{4}$$

It is not difficult to show that $\min(2a + b, p - b) \le \min(\frac{p-1}{2} + a, p - b)$ and thus Bzdęga's bound is never worse than Bachman's and in practice often strict inequality holds.

Note that if $q \equiv \pm 1 \pmod{p}$, then (3) implies that $A(pqr) \le (p+1)/2$, a result due to Beiter [1968] and, independently, Bloom [1968].

We remark that Bachman and Bzdęga define b as follows:

$$b = \min(b_1, p - b_1), \quad ab_1qr \equiv 1 \pmod{p}, \quad 0 < b_1 < p.$$

It is an easy exercise to see that our definition is equivalent to this one.

We will show that both (3) and (4) give rise to the same upper bound $f(q^*)$ for $M(p;q)$. Write $q^* \equiv j \pmod p$, $r^* \equiv k \pmod p$ with $1 \le j, k \le p-1$. Thus the right-hand sides of both (3) and (4) are functions of j and k, which we denote respectively by $\mathrm{GB}(j,k)$ and $\mathrm{BB}(j,k)$. We have

$$\mathrm{BB}(j,k) = \min(2a+b,\, p-b) \le \min\left(\frac{p-1}{2}+a,\, p-b\right) = \mathrm{GB}(j,k),$$

with $a = \min(j, k, p-j, p-k)$ and $b = \max(\min(j, p-j), \min(k, p-k))$.

Lemma 12. *Let* $1 \le j \le p-1$. *Denote* $\mathrm{GB}(j,j)$ *by* $f(j)$. *We have*

$$\max_{1\le k\le p-1} \mathrm{BB}(j,k) = \max_{1\le k\le p-1} \mathrm{GB}(j,k) = f(j),$$

with

$$f(j) = \begin{cases} \frac{1}{2}(p-1)+j & \text{if } j < p/4, \\ p-j & \text{if } p/4 < j \le \frac{1}{2}(p-1), \end{cases}$$

and $f(p-j) = f(j)$ *if* $j > \frac{1}{2}(p-1)$.

Proof. Since the problem is symmetric under replacing j by $p-j$, without loss of generality we may assume that $j \le \frac{1}{2}(p-1)$. If $j < p/4$, then

$$\mathrm{GB}(j,k) \le \frac{p-1}{2} + a \le \frac{p-1}{2} + j = \mathrm{GB}(j,j).$$

If $j > p/4$, then

$$\mathrm{GB}(j,k) \le p-b \le p-j = \mathrm{GB}(j,j).$$

Note that

$$\mathrm{GB}(j,j) = \begin{cases} \mathrm{BB}\big(j, \frac{1}{2}(p+1)-j\big) & \text{if } j < p/4, \\ \mathrm{BB}(j,j) & \text{if } j > p/4. \end{cases}$$

For example, if $j < p/4$, then the choice $q^* = j$, $r^* = \frac{1}{2}(p+1)-j$ leads to $a = j$ and $b = \frac{1}{2}(p+1)-j$ and hence

$$\mathrm{BB}\big(j, \tfrac{1}{2}(p+1)-j\big) = \min\big(\tfrac{1}{2}(p+1)+j,\, \tfrac{1}{2}(p-1)+j\big) = \mathrm{GB}(j,j).$$

Since $\mathrm{BB}(j,k) \le \mathrm{GB}(j,k) \le \mathrm{GB}(j,j)$ we are done. □

Theorem 13. *Let* $2 < p < q$. *Then* $M(p;q) \le f(q^*)$.

Proof. By (4) and the definition of $\mathrm{BB}(j,k)$ we have

$$M(p;q) \le \max_{1\le k\le p-1} \mathrm{BB}(q^*, k) = f(q^*),$$

completing the proof. □

Lemma 12 shows that using either (3) or (4), we cannot improve on the upper bound given in Theorem 13. Since

$$\max_{1\le j\le p-1} f(j) = p-1-\left[\frac{p}{4}\right] = \begin{cases} \frac{3}{4}(p-1) & \text{if } p\equiv 1 \pmod 4,\\ \frac{1}{4}(3p-1) & \text{if } p\equiv 3 \pmod 4,\end{cases}$$

we infer that

$$M(p) \le \max_{1\le j\le p-1}\ \max_{1\le k\le p-1} \mathrm{GB}(j,k) = \max_{1\le j\le p-1} f(j) < \tfrac{3}{4}p.$$

4. Earlier work on $M(p; q)$

Implicit in the literature are various results on $M(p; q)$ (although we are the first to explicitly study $M(p; q)$). Most of these are mentioned in the rest of this paper. Here we rewrite the main result of [Gallot and Moree 2009b] in terms of $M(p; q)$ and use it for $p = 11$, to deal with $q \equiv 4 \pmod{11}$, and $p = 13$, to deal with $q \equiv 5 \pmod{13}$.

Theorem 14. *Let $p \ge 11$ be a prime. Given any $1 \le \beta \le p-1$ we let β^* be the unique integer $1 \le \beta^* \le p-1$ with $\beta\beta^* \equiv 1 \pmod p$. Let $\mathcal{B}_-(p)$ be the set of integers satisfying*

$$1 \le \beta \le \frac{p-3}{2}, \quad p \le \beta + 2\beta^* + 1, \quad \beta > \beta^*.$$

Let $\mathcal{B}_+(p)$ be the set of integers satisfying

$$1 \le \beta \le \frac{p-3}{2}, \quad p \le \beta + \beta^*, \quad \beta \ge \beta^*/2.$$

Let $\mathcal{B}(p)$ be the union of these (disjoint) sets. As $(p-3)/2 \in \mathcal{B}(p)$, it is nonempty. Let $q \equiv \beta \pmod p$ be a prime satisfying $q > p$. Suppose that the inequality $q > q_-(p) := p(p-\beta^*)(p-\beta^*-2)/(2\beta)$ holds if $\beta \in \mathcal{B}_-(p)$ and

$$q > q_+(p) := \frac{p(p-1-\beta)}{\gamma(p-1-\beta)-p+1+2\beta},$$

with $\gamma = \min((p-\beta^*)/(p-\beta), (\beta^*-\beta)/\beta^*)$ if $\beta \in \mathcal{B}_+(p)$. Then

$$M(p; q) \ge p-\beta > \frac{p+1}{2}$$

and hence $M(p) \ge p - \min\{\mathcal{B}(p)\}$.

We have $\mathcal{B}(11) = \{4\}$, $\mathcal{B}(13) = \{5\}$, $\mathcal{B}(17) = \{7\}$ and $\mathcal{B}(19) = \{8\}$. In general one can show [Cobeli et al. $\ge$ 2011] using Kloosterman sum techniques that

$$\left| |\mathcal{B}(p)| - \frac{p}{16} \right| \le 24p^{3/4}\log p.$$

The lower bound for $M(p)$ resulting from this theorem, $p - \min\{\mathcal{B}(p)\}$, never exceeds $2p/3$ and this together with extensive numerical experimentation led in [Gallot and Moree 2009b] to the proposal of a corrected Beiter conjecture, now proved by Zhao and Zhang (Theorem 2).

Under the appropriate conditions on p and q, Theorem 14 says that $M(p; q) \geq p - \beta$, whereas Theorem 13 yields $M(p; q) \leq f(\beta^*)$. Thus studying the case $p - \beta = f(\beta^*)$ with $\beta \in \mathcal{B}(p)$, leads to a small subset of cases where $M(p; q)$ can be exactly computed using Theorem 14.

Theorem 15. *Let* $p \geq 13$ *with* $p \equiv 1 \pmod 4$ *be a prime. Let* x_0 *be the smallest positive integer such that* $x_0^2 + 1 \equiv 0 \pmod p$. *If* $x_0 > p/3$, $q \equiv x_0 \pmod p$ *and* $q \geq q_+(p)$ (*with* $\beta = x_0$), *then* $M(p; q) = p - x_0$.

Proof. Some easy computations show that if $p - \beta = f(\beta^*)$ and $\beta \in \mathcal{B}(p)$, we must have $\beta \in \mathcal{B}_+(p)$, $\frac{1}{2}(p-1) < \beta^* < \frac{3}{4}p$ and hence $f(\beta^*) = \beta^*$ and so

$$\beta \in \mathcal{B}_+(p), \quad 1 \leq \beta \leq \frac{p-3}{2}, \quad \beta + \beta^* = p, \quad \beta^* \leq 2\beta, \quad \frac{p-1}{2} < \beta^* < \frac{3}{4}p. \tag{5}$$

Note that $\beta + \beta^* = p$, $p \geq 13$, has a solution with $\beta < p/2$ if and only if $p \equiv 1 \pmod 4$ and $\beta = x_0$ (and hence $\beta^* = p - x_0$) with x_0 the smallest solution of $x_0^2 + 1 \equiv 0 \pmod p$. If $x_0 > p/3$, then $\beta = x_0$ satisfies (5). Since by assumption $q \geq q_+(p)$ and $q \equiv x_0 \pmod p$, we have $M(p; q) \geq p - x_0$ by Theorem 14. On the other hand, by Theorem 13, we have $M(p; q) \leq f(p - x_0) = f(x_0) = p - x_0$. □

Remark. The set of primes p satisfying $p \equiv 1 \pmod 4$ and $x_0 > p/3$ (which starts $\{13, 29, 53, 73, 89, 173, \dots\}$) has natural density $\frac{1}{6}$. This follows on taking $\alpha_2 = \frac{1}{2}$ and $\alpha_1 = \frac{1}{3}$ in the result from [Duke et al. 1995] that if f is a quadratic polynomial with complex roots and $0 \leq \alpha_1 < \alpha_2 \leq 1$ are prescribed real numbers, then as x tends to infinity,

$$\#\{(p, v) : p \leq x,\ f(v) \equiv 0 \pmod p,\ \alpha_1 \leq v/p < \alpha_2\} \sim (\alpha_2 - \alpha_1)\pi(x).$$

5. Computation of $M(3; q)$

Note that for all primes q and r with $1 < q < r$, there exists some unique $h \leq (q-1)/2$ and $k > 0$ such that $r = (kq+1)/h$ or $r = (kq-1)/h$. If $n \equiv 0 \pmod 3$ is ternary, then either $A(n) = 1$ or $A(n) = 2$ as $M(3) = 2$. The following result due to Sister Beiter [Beiter 1978] allows one to compute $A(n)$ in this case.

Theorem 16. *Let* $n \equiv 0 \pmod 3$ *be ternary.*

- *If* $h = 1$, *then* $A(n) = 1$ *if and only if* $k \equiv 0 \pmod 3$.
- *If* $h > 1$, *then* $A(n) = 1$ *if and only if one of the following conditions holds*:
  (a) $k \equiv 0 \pmod 3$ *and* $h + q \equiv 0 \pmod 3$.
  (b) $k \equiv 0 \pmod 3$ *and* $h + r \equiv 0 \pmod 3$.

We have seen that $M(3;5) = 2$. The next result extends this.

Theorem 17. *Let $q > 3$ be a prime. We have $M(3;q) = 2$.*

Proof. In case $q \equiv 1 \pmod 3$, then let r be a prime such that $r \equiv 1+q \pmod{3q}$. Since $(1+q, 3q) = 1$, Dirichlet's theorem says there are in fact infinitely many such primes. If $q \equiv 2 \pmod 3$, let r be a prime such that $r \equiv 1+2q \pmod{3q}$. Since $(1+2q, 3q) = 1$, there are infinitely many such primes. The prime r was chosen so as to ensure that $h = 1$ and $3 \nmid k$. Using Theorem 16 it then follows that $A(3qr) = 2$ and hence $M(3;q) = 2$. □

6. Kaplan's lemma reconsidered

Our main tool will be the following result of Kaplan, the proof of which uses the identity

$$\Phi_{pqr}(x) = (1+x^{pq}+x^{2pq}+\cdots)(1+x+\cdots+x^{p-1}-x^q-\cdots-x^{q+p-1})\Phi_{pq}(x^r).$$

Lemma 18 [Kaplan 2007]. *Let $2 < p < q < r$ be primes and $k \geq 0$ be an integer. Put*

$$b_i = \begin{cases} a_{pq}(i) & \text{if } ri \leq k, \\ 0 & \text{otherwise.} \end{cases}$$

We have

$$a_{pqr}(k) = \sum_{m=0}^{p-1} (b_{f(m)} - b_{f(m+q)}), \tag{6}$$

where $f(m)$ is the unique integer such that $f(m) \equiv r^{-1}(k-m) \pmod{pq}$ and $0 \leq f(m) < pq$.

(If we need to stress the k-dependence of $f(m)$, we will write $f_k(m)$ instead of $f(m)$, see, e.g., Lemma 22 and its proof.) This lemma reduces the computation of $a_{pqr}(k)$ to that of $a_{pq}(i)$ for various i. These binary cyclotomic polynomial coefficients are computed in the following lemma. For a proof see, e.g., [Lam and Leung 1996; Thangadurai 2000].

Lemma 19. *Let $p < q$ be odd primes. Let ρ and σ be the (unique) nonnegative integers for which $1 + pq = (\rho+1)p + (\sigma+1)q$. Let $0 \leq m < pq$. Then either $m = \alpha_1 p + \beta_1 q$ or $m = \alpha_1 p + \beta_1 q - pq$ with $0 \leq \alpha_1 \leq q-1$ the unique integer such that $\alpha_1 p \equiv m \pmod q$ and $0 \leq \beta_1 \leq p-1$ the unique integer such that $\beta_1 q \equiv m \pmod p$. The cyclotomic coefficient $a_{pq}(m)$ equals*

$$\begin{cases} 1 & \text{if } m = \alpha_1 p + \beta_1 q \text{ with } 0 \leq \alpha_1 \leq \rho,\ 0 \leq \beta_1 \leq \sigma, \\ -1 & \text{if } m = \alpha_1 p + \beta_1 q - pq \text{ with } \rho+1 \leq \alpha_1 \leq q-1,\ \sigma+1 \leq \beta_1 \leq p-1, \\ 0 & \text{otherwise.} \end{cases}$$

We say that $[m]_p = \alpha_1$ is the *p-part of m* and $[m]_q = \beta_1$ is the *q-part of m*. It is easy to see that

$$m = \begin{cases} [m]_p p + [m]_q q & \text{if } [m]_p \le \rho \text{ and } [m]_q \le \sigma; \\ [m]_p p + [m]_q q - pq & \text{if } [m]_p > \rho \text{ and } [m]_q > \sigma; \\ [m]_p p + [m]_q q - \delta_m pq & \text{otherwise,} \end{cases}$$

with $\delta_m \in \{0, 1\}$. Using this observation we find that, for $i < pq$,

$$b_i = \begin{cases} 1 & \text{if } [i]_p \le \rho, [i]_q \le \sigma \text{ and } [i]_p p + [i]_q q \le k/r; \\ -1 & \text{if } [i]_p > \rho, [i]_q > \sigma \text{ and } [i]_p p + [i]_q q - pq \le k/r; \\ 0 & \text{otherwise.} \end{cases}$$

Thus in order to evaluate $a_{pqr}(n)$ using Kaplan's lemma it suffices to compute $[f(m)]_p$, $[f(m)]_q$, and $[f(m+q)]_q$ (note that $[f(m)]_p = [f(m+q)]_p$).

For future reference we provide a version of Kaplan's lemma in which the computation of b_i has been made explicit, and thus is self-contained.

Lemma 20. *Let $2 < p < q < r$ be primes and let $k \ge 0$ be an integer. We put $\rho = [(p-1)(q-1)]_p$ and $\sigma = [(p-1)(q-1)]_q$. Furthermore, we put*

$$b_i = \begin{cases} 1 & \textit{if } [i]_p \le \rho, [i]_q \le \sigma \textit{ and } [i]_p p + [i]_q q \le k/r; \\ -1 & \textit{if } [i]_p > \rho, [i]_q > \sigma \textit{ and } [i]_p p + [i]_q q - pq \le k/r; \\ 0 & \textit{otherwise.} \end{cases}$$

We have

$$a_{pqr}(k) = \sum_{m=0}^{p-1} (b_{f(m)} - b_{f(m+q)}), \tag{7}$$

where $f(m)$ is the unique integer such that $f(m) \equiv r^{-1}(k-m) \pmod{pq}$ and $0 \le f(m) < pq$.

Note that if i and j have the same p-part, then $b_i b_j \ne -1$, that is b_i and b_j cannot be of opposite sign. From this it follows that $|b_{f(m)} - b_{f(m+q)}| \le 1$, and thus we infer from Kaplan's lemma that $|a_{pqr}(k)| \le p$ and hence $M(p) \le p$.

Using the mutual coprimality of p, q and r we arrive at the following trivial, but useful, lemma.

Lemma 21. *We have $\{[f(m)]_q : 0 \le m \le p-1\} = \{0, 1, 2, \ldots, p-1\}$ and $|\{[f(m)]_p : 0 \le m \le p-1\}| = p$. The same conclusions hold if we replace $[f(m)]_q$ and $[f(m)]_p$ by $[f(m+q)]_q$, respectively $[f(m+q)]_p$.*

Working with Kaplan's lemma one first computes $a_{pq}(f(m))$ and then $b_{f(m)}$. As a check on the correctness of the computations we note that the following identity should be satisfied.

Lemma 22. *We have*

$$\sum_{m=0}^{p-1} a_{pq}(f_k(m)) = \sum_{m=0}^{p-1} a_{pq}(f_k(m+q)).$$

Proof. Choose an integer $k_1 \equiv k \pmod{pq}$ such that $k_1 > pqr$. Then $a_{pqr}(k_1) = 0$. By Lemma 18 we find that

$$0 = a_{pqr}(k_1) = \sum_{m=0}^{p-1} \big(a_{pq}(f_{k_1}(m)) - a_{pq}(f_{k_1}(m+q))\big).$$

Since $f_k(m)$ only depends on the congruence class of k modulo pq, $f_{k_1}(m) = f_k(m)$ and the result follows. □

6.1. ***Working with Kaplan's lemma: examples.*** In this section we carry out some sample computations using Kaplan's lemma. For more involved examples the reader is referred to [Gallot and Moree 2009b].

We remark that the result that $a_n(k) = (p+1)/2$ in Lemma 23 is due to Herbert Möller [1971]. The proof we give here of this is rather different. The foundation for Möller's result is due to Emma Lehmer, who showed [1936] that

$$a_n\big(\tfrac{1}{2}(p-3)(qr+1)\big) = \tfrac{1}{2}(p-1)$$

with p, q, r and n satisfying the conditions of Lemma 23.

Lemma 23. *Let $p < q < r$ be primes satisfying*

$$p > 3, \quad q \equiv 2 \pmod{p}, \quad r \equiv \frac{p-1}{2} \pmod{p}, \quad r \equiv \frac{q-1}{2} \pmod{q}.$$

For $k = (p-1)(qr+1)/2$ we have $a_{pqr}(k) = (p+1)/2$.

Proof (taken from [Gallot and Moree 2009a]). Using that $q \equiv 2 \pmod{p}$, we infer from $1 + pq = (\rho+1)p + (\sigma+1)q$ that $\sigma = \frac{1}{2}(p-1)$ and $(\rho+1)p = 1 + \frac{1}{2}(p-1)q$ (and hence $\rho = (p-1)(q-2)/(2p)$). Invoking the Chinese remainder theorem one checks that

$$-r^{-1} \equiv 2 \equiv -\left(\frac{q-2}{p}\right) p + q \pmod{pq}. \tag{8}$$

Furthermore, writing $f(0)$ as a linear combination of p and q we see that

$$f(0) \equiv \frac{k}{r} \equiv \left(\frac{p-1}{2}\right) q + \frac{p-1}{2r} \equiv \left(\frac{p-1}{2}\right) q + 1 - p \equiv \rho p \pmod{pq}. \tag{9}$$

Since $f(m) \equiv f(0) - \frac{m}{r} \pmod{pq}$ we find using (8), (9) and the observation that $\rho - m(q-2)/p \ge 0$ for $0 \le m \le (p-1)/2$, that $[f(m)]_p = \rho - m(q-2)/p \le \rho$

and $[f(m)]_q = m \le \sigma$ for $0 \le m \le (p-1)/2$. Since $[f(m)]_p p + [f(m)]_q q = \rho p + 2m \le \rho p + p - 1 = [k/r]$, we deduce that $a_{pq}(f(m)) = b_{f(m)} = 1$ in this range; see also the following table:

m	$[f(m)]_p$	$[f(m)]_q$	$f(m)$	$a_{pq}(f(m))$	$b_{f(m)}$
0	ρ	0	ρp	1	1
1	$\rho-(q-2)/p$	1	$\rho p+2$	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	1	1
j	$\rho-j(q-2)/p$	j	$\rho p+2j$	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	1	1
$(p-1)/2$	0	$(p-1)/2$	$(p-1)q/2$	1	1

Note that $f(m) \equiv f(0) - m/r \equiv \rho p + 2m \pmod{pq}$, from which one easily infers that $f(m) = \rho p + 2m$ for $0 \le m \le p-1$ (as $\rho p + 2m \le \rho p + 2(p-1) < pq$). In the range $\frac{1}{2}(p+1) \le m \le p-1$ we have $f(m) \ge \rho p + p + 1 = (p-1)q/2 + 2 > k/r$, and hence $b_{f(m)} = 0$.

On noting that $f(m+q) \equiv f(m) - q/r \equiv f(m) + 2q \equiv \rho p + 2m + 2q \pmod{pq}$, one easily finds, for $0 \le m \le p-1$, that $f(m+q) = \rho p + 2m + 2q > k/r$ and hence $b_{f(m+q)} = 0$.

Invoking Kaplan's lemma one finds

$$a_{pqr}(k) = \sum_{m=0}^{p-1} b_{f(m)} - \sum_{m=0}^{p-1} b_{f(m+q)} = \frac{p+1}{2} - 0 = \frac{p+1}{2}. \qquad \square$$

Lemma 24. *Let* $3 < p < q < r$ *be primes satisfying*

$$q \equiv 1 \pmod{p}, \quad r^{-1} \equiv \frac{p+q}{2} \pmod{pq}.$$

For $k = (p-1)qr/2 - pr + 2$ *we have* $a_{pqr}(k) = -\min\left(\frac{q-1}{p} + 1, \frac{p+1}{2}\right)$.

Proof. Let $0 \le m \le p-1$. We have

$$\rho = \frac{(p-1)(q-1)}{p} \text{ and } \sigma = 0,$$

$$k \equiv 1 \pmod{p}, \quad k \equiv 0 \pmod{q}, \quad k \equiv 2 \pmod{r},$$

so that we can compute

$$[f(m)]_q \equiv q^{-1}r^{-1}(k-m) \equiv (1-m)/2 \pmod{p},$$
$$[f(m+q)]_q \equiv q^{-1}r^{-1}(k-m-q) \equiv -m/2 \pmod{p},$$
$$[f(m)]_p = [f(m+q)]_p \equiv p^{-1}r^{-1}(k-m) \equiv -m/2 \pmod{q}.$$

This leads to

$$[f(m)]_q = \begin{cases} (p+1-m)/2 & \text{for } m \text{ even,} \\ (2p+1-m)/2 & \text{for } m \text{ odd and } m \neq 1, \\ 0 & \text{for } m = 1, \end{cases}$$

$$[f(m+q)]_q = \begin{cases} (p-m)/2 & \text{for } m \text{ odd,} \\ (2p-m)/2 & \text{for } m \text{ even and } m \neq 0, \\ 0 & \text{for } m = 0, \end{cases}$$

$$[f(m)]_p = [f(m+q)]_p = \begin{cases} (q-m)/2 & \text{for } m \text{ odd,} \\ (2q-m)/2 & \text{for } m \text{ even and } m \neq 0, \\ 0 & \text{for } m = 0. \end{cases}$$

We consider four cases:

Case 1: $[f(m)]_p \le \rho$ and $[f(m)]_q \le \sigma$. In this case $m = 1$. Therefore

$$[f(m)]_p p + [f(m)]_q q = \frac{p(q-1)}{2} > \frac{k}{r}.$$

Case 2: $[f(m)]_p > \rho$ and $[f(m)]_q > \sigma$. This case only arises if m is even and $m \ge 2$. Then we have

$$\begin{aligned} [f(m)]_p p + [f(m)]_q q - pq &= \frac{2q-m}{2} p + \frac{p+1-m}{2} q - pq \\ &= \frac{q(p+1-m) - mp}{2} \le \frac{q(p-1)}{2} - p + \frac{2}{r} = \frac{k}{r}. \end{aligned}$$

However, not all even $m \ge 2$ satisfy $[f(m)]_p > \rho$. For this it is necessary that

$$\frac{2q-m}{2} > \frac{(p-1)(q-1)}{p}.$$

That means

$$\frac{m}{2} < \frac{q-1}{p} + 1$$

and since $0 < \frac{m}{2} \le \frac{p-1}{2}$ we have exactly $\min\left(\frac{q-1}{p}, \frac{p-1}{2}\right)$ different values of m.

Case 3: $[f(m+q)]_p \le \rho$ and $[f(m+q)]_q \le \sigma$. In this case we have $m = 0$. Therefore

$$[f(m+q)]_p p + [f(m+q)]_q q = 0 \le \frac{k}{r}.$$

Case 4: $[f(m+q)]_p > \rho$ and $[f(m+q)]_q > \sigma$. We must have $2|m$ and $m \ge 2$. We find

$$[f(m+q)]_p p + [f(m+q)]_q q - pq = \frac{2q-m}{2} p + \frac{2p-m}{2} q - pq > \frac{k}{r}.$$

This case analysis shows that (respectively)

$$\sum_{\substack{m=0\\ b_{f(m)}=1}}^{p-1} 1=0, \quad \sum_{\substack{m=0\\ b_{f(m)}=-1}}^{p-1} 1=\min\Bigl(\frac{q-1}{p},\frac{p-1}{2}\Bigr), \quad \sum_{\substack{m=0\\ b_{f(m+q)}=1}}^{p-1} 1=1, \quad \sum_{\substack{m=0\\ b_{f(m+q)}=-1}}^{p-1} 1=0.$$

Kaplan's lemma then yields

$$a_{pqr}(k)=\Bigl(0-\min\Bigl(\frac{q-1}{p},\frac{p-1}{2}\Bigr)\Bigr)-(1-0)=-\min\Bigl(\frac{q-1}{p}+1,\frac{p+1}{2}\Bigr). \quad \square$$

The next two lemmas are proved by application of Kaplan's lemma; see [Gallot et al. 2010] for details.

Lemma 25. *Let* $3 < p < q < r$ *be primes satisfying*

$$q \equiv -2 (\text{mod } p), \quad r^{-1} \equiv p-2 \ (\text{mod } pq) \text{ and } q > p^2/2.$$

For $k=\frac{p+1}{2}(1+r(2-p+q))+r+q-rq$ *we have* $a_{pqr}(k)=-(p+1)/2$.

Remark. Numerical experimentation suggests that with this choice of k, a condition of the form $q > p^2c_1$, with c_1 some absolute positive constant, is unavoidable.

Lemma 26. *Let* $3 < p < q < r$ *be primes satisfying*

$$q \equiv -1 \ (\text{mod } p), \quad r^{-1} \equiv \frac{p+q}{2} \ (\text{mod } pq) \text{ and } q \geq p^2-2p.$$

For $k=p(q-1)r/2-rq+p-1$ *we have* $a_{pqr}(k)=-(p+1)/2$.

Proof of Proposition 9. The first assertion follows by Theorem 6, so assume $p \geq 11$. We will argue by contradiction. So suppose that $p \nmid \mathfrak{f}_p$. Put $\beta=(p-3)/2$. By the Chinese remainder theorem and Dirichlet's theorem there are infinitely many primes q_1 such that $q_1 \equiv 2 \ (\text{mod } p)$ and $q_1 \equiv 1 \ (\text{mod } \mathfrak{f}_p)$. Further, there are infinitely many primes q_2 such that $q_2 \equiv \beta \ (\text{mod } p)$ and $q_2 \equiv 1 \ (\text{mod } \mathfrak{f}_p)$. By the definition of $\mathfrak{f}_p$ there exists an integer c such that $M(p;q)=c$ for all $q \equiv 1 \ (\text{mod } \mathfrak{f}_p)$ that are large enough. However, by Lemma 23 we have $M(p;q_1)=(p+1)/2$ and by Theorem 14 (note that $\beta \in \mathcal{B}(p)$) we have $M(p;q_2) > (p+1)/2$ for all q_2 large enough. This contradiction shows that $p \nmid \mathfrak{f}_p$. $\square$

The results from this section together with those from Section 3 allow one to establish the following theorem. In Section 10 we will discuss the sharpness of the lower bounds for q.

Theorem 27. *Let* $2 < p < q$ *be primes.*

(a) *If* $q \equiv 2 \ (\text{mod } p)$, *then* $M(p;q)=(p+1)/2$.

(b) *If* $q \equiv -2 \ (\text{mod } p)$ *and* $q > p^2/2$, *then* $M(p;q)=(p+1)/2$.

(c) *If* $q \equiv 1 \ (\text{mod } p)$ *and* $q \geq (p-1)p/2+1$, *then* $M(p;q)=(p+1)/2$.

(d) *If* $q \equiv -1 \ (\text{mod } p)$ *and* $q \geq p^2-2p$, *then* $M(p;q)=(p+1)/2$.

Proof. By Theorem 17 we have $M(3; q) = 2 = (3 + 1)/2$, so assume $p > 3$.

(a) We have $M(p; q) \geq (p + 1)/2$ by Lemma 23, and $M(p; q) \leq f(2^*) = f((p + 1)/2) = (p + 1)/2$ by Theorem 13.

(b)+(c)+(d) Similar to that of part (a). Note that $f((-2)^*) = f((p - 1)/2) = (p + 1)/2$ and $f(1) = f(p - 1) = (p + 1)/2$. □

Theorem 28. *Let $q > 5$ be a prime. Then $M(5; q) = 3$.*

Proof. The proof is most compactly given in a table:

$\bar{q}$	q_0	$M(5; q)$	result
1	11	3	Theorem 27(c)
2	7	3	Theorem 27(a)
3	13	3	Theorem 27(b)
4	19	3	Theorem 27(d)

Interpretation: the third row, for example, says that for $q \equiv 3 \pmod 5$, $q \geq 13$, we have $M(5; q) = 3$ by Theorem 27(b). □

7. Computation of $M(7; q)$

Theorem 27, together with the next two lemmas (again proved by application of Kaplan's lemma), allows one to compute $M(7; q)$. These lemmas concern the computation of $M(p; q)$ with $q \equiv (p \pm 1)/2 \pmod p$.

Lemma 29. *Let $p \geq 5$ be a prime. Let $q \geq \max(3p, p(p + 1)/4)$ be a prime satisfying $q \equiv (p - 1)/2 \pmod p$. Let $r > q$ be a prime satisfying*

$$r^{-1} \equiv \frac{p+1}{2} \pmod p, \quad r^{-1} \equiv p \pmod q.$$

For $k = p - 1 + r(1 + q(p - 1)/2 - p(p + 1)/2)$ we have $a_{pqr}(k) = (p + 1)/2$.

Lemma 30. *Let $p \geq 5$ be a prime. Let $q \geq \max(3p, p(p - 1)/4 + 1)$ be a prime satisfying $q \equiv (p + 1)/2 \pmod p$. Let $r > q$ be a prime satisfying*

$$r^{-1} \equiv \frac{p-1}{2} \pmod p, \quad r^{-1} \equiv p \pmod q.$$

For $k = q + p - 1 + r(q(p - 1)/2 - p(p + 1)/2)$ we have $a_{pqr}(k) = (p + 1)/2$.

Theorem 31.

(a) *If $q \geq \max(3p, p(p+1)/4)$ is a prime satisfying $q \equiv (p-1)/2 \pmod p$, then $(p + 1)/2 \leq M(p; q) \leq (p + 3)/2$.*

(b) *If $q \geq \max(3p, p(p - 1)/4 + 1)$ is a prime satisfying $q \equiv (p + 1)/2 \pmod p$, then $(p + 1)/2 \leq M(p; q) \leq (p + 3)/2$.*

Proof. This follows on noting that

$$f\left(\left(\frac{p+1}{2}\right)^*\right) = f(2) = \frac{p+3}{2} = f(p-2) = f\left(\left(\frac{p-1}{2}\right)^*\right),$$

and combining Lemmas 29 and 30 with Theorem 13. □

Theorem 32. *We have* $M(7;\,11) = 4$, $M(7;\,13) = 3$ *and for* $q \geq 17$ *a prime*, $M(7;\,q) = 4$.

Proof. Again we encode the proof in a table:

$\bar{q}$	q_0	$M(7;\,q)$	result
1	29	4	Theorem 27(c)
2	23	4	Theorem 27(a)
3	31	4	Theorem 31(a)*
4	53	4	Theorem 31(b)*
5	47	4	Theorem 27(b)
6	41	4	Theorem 27(d)

For the entries marked with asterisks we also need the fact that $M(7) \leq 4$ (see just before Theorem 2). Since $M(7;\,11) = M(7;\,17) = M(7;\,19) = 4$ and $M(7;\,13) = 3$ (the only cases not covered in the table), the proof is completed. □

Proof of Theorem 6. Combine Theorems 17, 28 and 32. □

8. Computation of $M(11;\,q)$

We have $M(11;\,q) \leq M(11) = 7$ (by Theorem 2 and Table 1). Moreover:

Theorem 33 [Gallot and Moree 2009b]. *Let* $q < r$ *be primes with* $q \equiv 4 \pmod{11}$ *and* $r \equiv -3 \pmod{11}$. *Let* $1 \leq \alpha \leq q-1$ *be the unique integer such that* $11r\alpha \equiv 1 \pmod{q}$. *Suppose that* $q/33 < \alpha \leq (3q-1)/77$. *Then* $a_{11qr}(10+(6q-77\alpha)r) = -7$.

Lemma 34. *Let q be a prime such that* $q \equiv 4 \pmod{11}$. *For* $q > 37$, $M(11;\,q) = 7$, *and* $M(11;\,37) = 6$.

Proof. By computation one finds that $M(11;\,37) = 6$. Now assume $q > 37$. Notice that it is enough to show that $M(11;\,q) \geq 7$. For $q \geq 191$ the interval $I(q) := (q/33,\,(3q-1)/77]$ has length exceeding 1 and so contains at least one integer α_1. Then by the Chinese remainder theorem and Dirichlet's theorem we can find a prime r_1 such that both $r_1 \equiv -3 \pmod{11}$ and $11r_1\alpha_1 \equiv 1 \pmod{q}$. Then we invoke Theorem 33 with $r = r_1$ and $\alpha = \alpha_1$. It remains to deal with the primes 59 and 103. One checks that both intervals $I(59)$ and $I(103)$ contain an integer and so we can proceed as in the case $q \geq 191$ to conclude the proof. □

Lemma 35. *Let* $p = 11$.

(a) *For* ≥ 133, $q \equiv 3 \pmod{11}$, $r^{-1} \equiv \frac{q-19}{2} \pmod{pq}$ *and* $k = q + 7r\frac{(q-19)}{2}$ *we have* $a_{pqr}(k) = 7$.

(b) *For* $q \equiv 7 \pmod{11}$, $r^{-1} \equiv \frac{q+7}{2} \pmod{pq}$ *and* $k = 6qr+4$ *we have* $a_{pqr}(k) = 7$.

(c) *For* $q \equiv 8 \pmod{11}$, $r^{-1} \equiv \frac{q-3}{2} \pmod{pq}$ *and* $k = 6qr+4$ *we have* $a_{pqr}(k) = 7$.

The proof is an application of Kaplan's lemma.

Theorem 36. *For* $q \geq 13$ *we have*

$q \pmod{11}$	1	2	3	4	5	6	7	8	9	10
$M(11;q)$	6	6	7	7	6,7	6,7	7	7	6	6

except when $q \in \{17, 23, 37, 43, 47\}$. *We have* $M(11; 17) = 5$, $M(11; 23) = 3$, $M(11; 37) = 6$, $M(11; 43) = 5$ *and* $M(11; 47) = 6$.

Remarks. (1) If $q \equiv \pm 5 \pmod{11}$ and $q \geq 61$, then $M(p, q) \in \{6, 7\}$. We believe that $M(p; q) = 6$.

(2) By Corollary 41 and 42 following Theorem 40, one infers that $M(11; 17) \leq 5$, $M(11; 23) \leq 3$ and $M(11; 43) \leq 5$.

Proof of 36.

$\bar{q}$	q_0	$M(11;q)$	result
1	67	6	Theorem 27(c)
2	13	6	Theorem 27(a)
3	157	7	Lemma 35(a)*
4	59	7	Lemma 34
5	71	6,7	Theorem 31(a)*
6	61	6,7	Theorem 31(b)*
7	29	7	Lemma 35(b)*
8	19	7	Lemma 35(c)*
9	97	6	Theorem 27(b)
10	109	6	Theorem 27(d)

Here the asterisks indicate that we need the fact that $M(11) = 7$. The proof is completed by directly computing the values of $M(p; q)$ not covered by the table. □

9. Computation for $p = 19$

By Theorem 2 we have $M(19) \leq 2 \cdot 19/3$ and hence $M(19) \leq 12$. By Theorem 14 we find that $M(19; q) \geq 11$ for every $q \equiv 8 \pmod{19}$ and $q \geq 179$ and hence

$M(19) \geq 11$. Since $A(19 \cdot 53 \cdot 859) = 12$, it follows that $M(19) = 12$. The next result even shows that $M(19; q) = M(19)$ for a positive fraction of the primes.

Theorem 37. *We have* $M(19) = 12$. *Moreover*, $M(19, q) = 12$ *if* $q \equiv \pm 4 \pmod{19}$, *with* $q > 23$. *Furthermore*, $M(19; 23) = 11$.

The proof is an almost direct consequence of the following lemma, itself proved by applying Kaplan's lemma.

Lemma 38. *Put* $p = 19$ *and let* $q \equiv \pm 4 \pmod{19}$ *be a prime. Suppose there exists an integer a satysifying*

$$qa \equiv -1 \pmod 3 \text{ and } \frac{q}{6p} < a \leq \frac{5q-18}{6p}. \tag{10}$$

Let $r > q$ *be a prime satisfying* $r(q-ap) \equiv 3 \pmod{pq}$. *Then* $a_{pqr}(7qr+q) = -12$, *if* $q \equiv -4 \pmod{19}$, *and* $a_{19qr}(7qr+r) = -12$ *if* $q \equiv 4 \pmod{19}$.

Proof of Theorem 37. For $q > 90$ the interval in (10) is of length > 3 and so contains an integer a satisfying $qa \equiv -1 \pmod 3$. It remains to deal with $q \in \{23, 53, 61\}$. Computation shows that $M(19; 23) = 11$. For $q = 53$ and $q = 61$ one finds an integer a satisfying condition (10). □

Proof of Theorem 5. By Theorem 14 and Dirichlet's theorem the claim follows for $p = 13$. Using Lemmas 34 and 35 the result follows for $p = 11$. On invoking Theorems 6 and 37, the proof is then completed. □

10. Small values of $M(p; q)$

Typically if $M(p; q)$ is constant for all q large enough with $q \equiv a \pmod d$, then $M(p; q)$ assumes a smaller value for some small q in this progression. A (partial) explanation of this phenomenon is provided in this section. We will show that if $ap + bq = 1$ with a and b small in absolute value, then $M(p; q)$ is small. On the other hand we will show that $M(p; q)$ cannot be truly small.

Proposition 39. *Let* $2 < p < q$ *be odd primes. Then* $M(p; q) \geq 2$.

Proof. We say $\Phi_n(x)$ is flat if $A(n) = 1$. ChunGang Ji [2010] proved that if $p < q < r$ are odd prime and $2r \equiv \pm 1 \pmod{pq}$, then $\Phi_{pqr}(x)$ is flat if and only if $p = 3$ and $q \equiv 1 \pmod 3$. It follows that $M(p; q) \geq 2$ for $p > 3$. Now invoke Theorem 17 to deal with the case $p = 3$. □

Theorem 40. *Let* $2 < p < q$ *be odd primes and* ρ *and* σ *be the* (*unique*) *nonnegative integers for which* $1 + pq = (\rho+1)p + (\sigma+1)q$. *Then*

$$M(p; q) \leq \begin{cases} p + \rho - \sigma & \text{if } \rho \leq \sigma, \\ q + \sigma - \rho & \text{if } \rho > \sigma. \end{cases}$$

Corollary 41. *Let h, k be integers with $k > h$ and $q = (kp - 1)/h$ a prime. If $p \geq k + h$, then $M(p; q) \leq k + h$.*

Corollary 42. *Let h, k be integers with $k > h$ and $q = (kp + 1)/h$ a prime. If $p > h$ and $q > k + h$, then $M(p; q) \leq k + h$.*

Proof of Theorem 40. Let us assume that $\rho \leq \sigma$, the other case being similar. Using Lemma 21 and Lemma 19 we infer that the number of $0 \leq m \leq p-1$ with $b_{f(m)} = 1$ is at most $\rho+1$. Likewise the number of m with $b_{f(m+q)} = -1$ is at most $p-1-\sigma$. By Kaplan's lemma it then follows that $a_{pqr}(k) \leq \rho+1+(p-1-\sigma) = p+\rho-\sigma$. Since the number of $0 \leq m \leq p-1$ with $b_{f(m)} = -1$ is at most $p-1-\sigma$ and the number of m with $b_{f(m+q)} = 1$ is at most $\rho+1$, we infer that $a_{pqr}(k) \geq -(p+\rho-\sigma)$ and hence the result is proved. □

Theorem 43. *Let $q \equiv 1 \pmod{p}$. Then*

$$M(p; q) = \min\left(\frac{q-1}{p} + 1, \frac{p+1}{2}\right).$$

Proof. For $p = 3$ the result follows by Theorem 17, so assume $p \geq 5$. Sister Beiter [Beiter 1968], and independently Bloom [Bloom 1968], proved that $M(p; q) \leq (p + 1)/2$ if $q \equiv \pm 1 \pmod{p}$ (alternatively we invoke Theorem 13). By Corollary 42 we have $M(p; q) \leq (q - 1)/p + 1$. By Lemma 24 the proof is then completed. □

Numerical experiments suggest that in Theorem 27(b) the condition $q > p^2/2$ can perhaps be dropped. By Theorem 43 the condition $q \geq (p-1)p/2+1$ in part (c) is optimal. In (d) we need $q \geq (p-1)p/2-1$; otherwise $M(p; q) < (p+1)/2$ by Corollary 41.

Lemma 44. *Let $p \geq 7$ be a prime such that $q = 2p - 1$ is also a prime. Let $r > q$ be a prime such that $(p + q)r \equiv -2 \pmod{pq}$. Put $k = rq(p - 1)/2 + 2p - pq$. Then $a_{pqr}(k) = 3$.*

The proof is an application of Kaplan's lemma.

Proof of Theorem 7. On combining Lemma 44 with Corollary 41, one deduces that $M(p; 2p - 1) = 3$ if $p \geq 5$ and $2p - 1$ is a prime. □

11. Conjectures, questions, problems

The open problem that we think is the most interesting is Conjecture 8. If one could prove it and obtain an effective upper bound for the ternary conductor $\mathfrak{f}_p$ (say $16p$) and an effective upper bound for the minimal ternary prime (say p^3), one would have a finite procedure to compute $M(p)$.

Problem 45. Bachman [2010] introduced inclusion-exclusion polynomials. These polynomials generalize the ternary cyclotomic polynomials. Study $M(p;q)$ in this setting (here p and q can be any coprime natural numbers), cf. Section 2 where we denoted this function by $M'(p;q)$. For example, using [Bachman 2010, Theorem 3] by an argument similar to that given in Proposition 3 it is easily seen that there is a finite procedure to compute $M'(p;q)$.

Problem 46. The analogue of $M(p;q)$ for inverse cyclotomic polynomials can be defined [Moree 2009]. Study it.

Question 47. Can one compute the average value of $M(p;q)$, that is does the limit

$$\lim_{x\to\infty}\frac{1}{\pi(x)}\sum_{p<q\le x} M(p;q)$$

exist and if yes, what is its value?

Question 48. Is Theorem 5 still true if we put $\delta(13)=1/3$ and cross out the words “a subset having”?

Question 49. If $q>p$ is prime and $q\equiv -2 \pmod{p}$, then do we have $M(p;q)=(p+1)/2$?

Question 50. Suppose that $p>11$ is a prime.
If $6p-1$ is prime, then do we have $M(p,6p-1)=7$?
If $(5p-1)/2$ is prime, then do we have $M(p,(5p-1)/2)=7$?
If $(5p+1)/2$ is prime then do we have $M(p,(5p+1)/2)=7$?
Find more similar results.

Question 51. Given an integer $k\ge 1$, does there exist $p_0(k)$ and a function $q_k(p)$ such that if $q\equiv 2/(2k+1)(\mathrm{mod}\ p)$, $q\ge q_k(p)$ and $p\ge p_0(k)$, then $M(p;q)=(p+2k+1)/2$?

Question 52. Is it true that $M(11;q)=6$ for all large enough q satisfying $q\equiv \pm 5 \pmod{6}$? If so one can finish the computation of $M(11;q)$.

Question 53. Is it true that for q sufficiently large the values of $M(13;q)$, $M(17;q)$, $M(19;q)$ and $M(23;q)$ are given by Table 2 on the next page?

The next question was raised by the referee of this paper.

Question 54. Suppose that for all sufficiently large primes $q\equiv q_0 \pmod{\mathfrak{f}_p}$ we have $M(p;q)<M(p)$. Is it possible to prove that $M(p;q)<M(p)$ for every prime $q\equiv q_0 \pmod{\mathfrak{f}_p}$?

Question 55. For a given prime p, let $m(p)$ denote $\liminf M(p;q)$, with $q>p$. Determine $m(p)$. Is it true that $\lim_{p\to\infty} m(p)/p=c$ for some constant $c>0$?

q (mod 13)	1	2	3	4	5	6	7	8	9	10	11	12
$M(13;q)$	7	7	7	8	8	7	7	8	8	7	7	7

q (mod 17)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$M(17;q)$	9	9	9	10	10	9	10	9	9	10	9	10	10	9	9	9

q (mod 19)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$M(19;q)$	10	10	10	12	11	9	11	11	10	10	11	11	9	11	12	10	10	10

q (mod 23)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
$M(23;q)$	12	12	12	14	14	11	13	11	14	13	12	12	13	14	11	13	11	14	14	12	12	12

Table 2. Conjectural values of $M(13;q)$, $M(17;q)$, $M(19;q)$ and $M(23;q)$ (for q large). See Question 53.

By Proposition 39 we have $m(p) \geq 2$ for $p > 2$. Note that the results in this paper imply that $m(p) = (p+1)/2$ for $2 < p \leq 11$. If the answer to Question 53 is yes, then $m(p) = (p+1)/2$ for $2 < p \leq 17$ and $m(p) = (p-1)/2$ for $19 \leq p \leq 23$. (The issue of lower bounds for $M(p;q)$ was raised by the referee.)

Acknowledgement

Moree thanks N. Baghina, C. Budde, B. Jüttner and D. Sullivan, interns at the Max-Planck-Institut für Mathematik in June 2008, for their computer assistance in filling in the tables used in the proofs of Lemma 38 ($p = 19$). For these tables see [Gallot et al. 2010]. However, the bulk of the paper was written whilst Wilms was during two months in 2010 an intern at MPIM under Moree's guidance. Wilms thanks the MPIM for the possibility to do an internship and for the nice research atmosphere. He also thanks Moree for his mentoring and for having a sympathetic ear for any questions.

Finally thanks are due to G. Bachman for some helpful remarks, and the referee who spent quite a bit of time writing a very extensive report that led to many improvements over the original submission.

References

[Bachman 2003] G. Bachman, "On the coefficients of ternary cyclotomic polynomials", *J. Number Theory* **100**:1 (2003), 104–116. MR 2004a:11020 Zbl 1023.11010

[Bachman 2004] G. Bachman, "Ternary cyclotomic polynomials with an optimally large set of coefficients", *Proc. Amer. Math. Soc.* **132**:7 (2004), 1943–1950. MR 2005c:11157 Zbl 1050.11027

[Bachman 2010] G. Bachman, "On ternary inclusion-exclusion polynomials", *Integers* **10**:5 (2010), 623–638. MR 2798626 Zbl 1213.11056 arXiv 1006.0518

[Bachman and Moree 2011] G. Bachman and P. Moree, "On a class of ternary inclusion-exclusion polynomials", *Integers* **11**:1 (2011), 77–91. MR 2798664 Zbl 1226.11032

[Bang 1895] A. S. Bang, "Om ligningen $\varphi_n(x) = 0$", *Nyt Tidsskrift for Matematik* (*B*) **6** (1895), 6–12. JFM 26.0121.01

[Beiter 1968] M. Beiter, "Magnitude of the coefficients of the cyclotomic polynomial $F_{pqr}(x)$", *Amer. Math. Monthly* **75**:4 (1968), 370–372. MR 37 #2670 Zbl 0157.08804

[Beiter 1971] M. Beiter, "Magnitude of the coefficients of the cyclotomic polynomial F_{pqr}, II", *Duke Math. J.* **38**:3 (1971), 591–594. MR 43 #6152 Zbl 0221.10018

[Beiter 1978] M. Beiter, "Coefficients of the cyclotomic polynomial $F_{3qr}(x)$", *Fibonacci Quart.* **16**:4 (1978), 302–306. MR 80a:10026 Zbl 0396.10006

[Bloom 1968] D. M. Bloom, "On the coefficients of the cyclotomic polynomials", *Amer. Math. Monthly* **75**:4 (1968), 372–377. MR 37 #2671 Zbl 0157.08901

[Bzdęga 2010] B. Bzdęga, "Bounds on ternary cyclotomic coefficients", *Acta Arith.* **144**:1 (2010), 5–16. MR 2011g:11204 Zbl 05834820

[Cobeli et al. $\geq$ 2011] C. Cobeli, Y. Gallot, P. Moree, and A. Zaharescu, "Distribution of modular inverses and large cyclotomic coefficients". In preparation.

[Duke et al. 1995] W. Duke, J. B. Friedlander, and H. Iwaniec, "Equidistribution of roots of a quadratic congruence to prime moduli", *Ann. of Math.* (2) **141**:2 (1995), 423–441. MR 95k:11124 Zbl 0840.11003

[Fintzen 2011] J. Fintzen, "Cyclotomic polynomial coefficients $a(n, k)$ with n and k in prescribed residue classes", *J. Number Theory* **131**:10 (2011), 1852–1863. MR 2811553 Zbl 05931159

[Gallot and Moree 2009a] Y. Gallot and P. Moree, "Neighboring ternary cyclotomic coefficients differ by at most one", *J. Ramanujan Math. Soc.* **24**:3 (2009), 235–248. MR 2010j:11045 Zbl 1205.11033 arXiv 0810.5496

[Gallot and Moree 2009b] Y. Gallot and P. Moree, "Ternary cyclotomic polynomials having a large coefficient", *J. Reine Angew. Math.* **632** (2009), 105–125. MR 2010g:11187 Zbl 1230.11030

[Gallot et al. 2010] Y. Gallot, P. Moree, and R. Wilms, "The family of ternary cyclotomic polynomials with one free prime", preprint 2010-11, Max-Planck-Institut für Mathematik, Bonn, 2010, available at http://www.mpim-bonn.mpg.de/node/263. A longer version of this paper (32 pages) with some proofs given in greater detail.

[Ji 2010] C. Ji, "A specific family of cyclotomic polynomials of order three", *Sci. China Math.* **53**:9 (2010), 2269–2274. MR 2011h:11025 Zbl 1229.11048

[Kaplan 2007] N. Kaplan, "Flat cyclotomic polynomials of order three", *J. Number Theory* **127**:1 (2007), 118–126. MR 2008k:11031 Zbl 1171.11015

[Lam and Leung 1996] T. Y. Lam and K. H. Leung, "On the cyclotomic polynomial $\Phi_{pq}(X)$", *Amer. Math. Monthly* **103**:7 (1996), 562–564. MR 97h:11150 Zbl 0868.11016

[Lehmer 1936] E. Lehmer, "On the magnitude of the coefficients of the cyclotomic polynomial", *Bull. Amer. Math. Soc.* **42**:6 (1936), 389–392. MR 1563307 Zbl 0014.39203

[Möller 1971] H. Möller, "Über die Koeffizienten des n-ten Kreisteilungspolynoms", *Math. Z.* **119**:1 (1971), 33–40. MR 43 #148 Zbl 0196.07201

[Moree 2009] P. Moree, “Inverse cyclotomic polynomials”, *J. Number Theory* **129**:3 (2009), 667–680. MR 2009k:11199 Zbl 1220.11037

[Thangadurai 2000] R. Thangadurai, “On the coefficients of cyclotomic polynomials”, pp. 311–322 in *Cyclotomic fields and related topics* (Pune, 1999), edited by S. D. Adhikari et al., Bhaskaracharya Pratishthana, Pune, 2000. MR 2001k:11213 Zbl 1044.11093

[Zhao and Zhang 2009] J. Zhao and X. Zhang, “A proof of the corrected Beiter conjecture”, preprint, 2009. arXiv 0910.2770

[Zhao and Zhang 2010] J. Zhao and X. Zhang, “Coefficients of ternary cyclotomic polynomials”, *J. Number Theory* **130**:10 (2010), 2223–2237. MR 2011d:11057 Zbl 05798167

Received: 2010-07-21 Revised: 2011-07-19 Accepted: 2011-08-15

galloty@orange.fr *12 bis rue Perrey, 31400 Toulouse, France*

moree@mpim-bonn.mpg.de *Max-Planck-Institut für Mathematik, Vivatsgasse 7,*
D-53111 Bonn, Germany

robert.wilms@rub.de *Sterbeckerstrasse 21, D-58579 Schalksmühle, Germany*

INVOLVE 4:4(2011)

Preimages of quadratic dynamical systems

Benjamin Hutz, Trevor Hyde and Benjamin Krause

(Communicated by Bjorn Poonen)

For a quadratic polynomial with rational coefficients, we consider the problem of bounding the number of rational points that eventually land at a given constant after iteration, called preimages of the constant. It was shown by Faber, Hutz, Ingram, Jones, Manes, Tucker, and Zieve (2009) that the number of rational preimages is bounded as one varies the polynomial. Explicit bounds on the number of preimages of zero and -1 were addressed in subsequent articles. This article addresses explicit bounds on the number of preimages of any algebraic number for quadratic dynamical systems and provides insight into the geometric surfaces parameterizing such preimages.

1. Introduction

Fix an algebraic number field K and a number $c \in K$ and define an endomorphism of the affine line by

$$f_c : \mathbb{A}^1_K \to \mathbb{A}^1_K, \qquad f_c(x) = x^2 + c.$$

If we define f_c^N to be the N-fold composition of the morphism f_c, and f_c^{-N} to be the inverse image of a in $\mathbb{A}^1_K$ under f_c^N, then for $a \in \mathbb{A}^1(K)$, the set of *rational iterated preimages of a* is given by

$$\bigcup_{N \geq 1} f_c^{-N}(a)(K) = \{x_0 \in \mathbb{A}^1(K) : f_c^N(x_0) = a \text{ for some } N \geq 1\}.$$

Heuristically, finding iterated preimages amounts to solving progressively more complicated polynomial equations, so K-rational solutions should be a rarity. The situation becomes more interesting as we vary c, which has the effect of varying the morphism f_c.

MSC2010: primary 37P05, 14G05; secondary 37F10.

Keywords: quadratic dynamical systems, arithmetic geometry, preimage, rational points, uniform bound.

The authors wish to thank David Cox for patiently answering many geometric questions, the Amherst Computing Cluster for providing CPU time, and the many helpful comments and suggestions from the referee.

Definition 1.1. Define

$$\kappa(a) = \sup_{c\in K} \#\left\{ \bigcup_{N\geq 1} f_c^{-N}(a)(K) \right\}.$$

A special case of the main theorem in [Faber et al. 2009] shows that $\kappa(a)$ is finite, but does not give an explicit bound. Note that it is easy to construct a pair (a, c) with arbitrarily many rational preimages simply by fixing c and taking $a = f_c(N)(0)$. The fact that $\kappa(a)$ is finite shows that, for a given a, such c values are rarely defined over the same field.

When needed for clarity, we include the field K in the notation as $\kappa(a, K)$. In this article, we focus on a weaker notion $\bar{\kappa}(a)$ that bounds the "typical" number of rational preimages.

Definition 1.2. Define

$$\bar{\kappa}(a, K) = \limsup_{c\in K} \#\left\{ \bigcup_{N\geq 1} f_c^{-N}(a)(K) \right\}.$$

In essence $\bar{\kappa}(a)$ differs from $\kappa(a)$ by excluding at most finitely many c values from consideration, thus, $\bar{\kappa}(a) \leq \kappa(a)$.

The cases of $a = 0$ and $a = -1$ were studied in [Faber et al. 2011; Hyde 2010], respectively, and it was shown that

$$\bar{\kappa}(0, \mathbb{Q}) = \bar{\kappa}(-1, \mathbb{Q}) = 6.$$

In the first of these papers, a significant amount of effort went into the more difficult task of showing that $\kappa(0, \mathbb{Q}) = 6$, assuming some standard conjectures. This article addresses the situation from the more general setting of allowing a to vary and examining the "preimage surfaces" instead of "preimage curves." We also allow arbitrary number fields K. Our main result is the following theorem.

Theorem 1.3. *For $a \in \overline{\mathbb{Q}}$ and for any fixed algebraic number field K we have*

$$\bar{\kappa}(a, K) = \begin{cases} 10 & \textit{if } a = -\frac{1}{4}, \\ 6 \textit{ or } 8 & \textit{if } a \textit{ is one of the three third critical values}, \\ 4 & \textit{if } a \in S \cap K, \\ 6 & \textit{otherwise}. \end{cases}$$

The set S is the finite set of a values (in $\overline{\mathbb{Q}}$) where the elliptic surface with two rational first preimages and four rational second preimages and the elliptic surface with two rational first preimages, (at least) two rational second preimages, and (at least) two rational third preimages both have specialization with rank zero at a.

The elliptic surface parameterizing values of a and c with two rational first preimages, (at least) two rational second preimages, and (at least) two rational third preimages has generic rank two (Theorem 3.3). Thus, finding the set of a values where the corresponding specialization is an elliptic curve of rank zero is a generalization of the problem studied by Masser and Zannier [2008]. The same authors have shown that such sets are finite [Masser and Zannier 2012], implying the set S is finite. The critical values are defined in Definition 2.1.

The organization of the article is as follows. In Section 3 we examine the lower bound for $\bar{\kappa}(a)$ by finding the generic rank over $\mathbb{Q}$ of the elliptic surfaces corresponding to arrangements of 6 preimages. In Section 4 we examine the upper bound on $\bar{\kappa}(a)$ by showing that all arrangements of $2N$ preimages for some N correspond to curves of genus greater than 1. In Section 5 we prove Theorem 1.3. In Section 6 we prove some additional properties of the preimage surfaces that are tangential to the proof of Theorem 1.3, yet still of interest. First we parameterize the possible torsion subgroups of the elliptic surface corresponding to two rational first preimages and four rational second preimages. Then, starting on page 362, we examine exceptional pairs (a, c) that are excluded by considering $\bar{\kappa}(a)$ instead of $\kappa(a)$.

We present these results for two reasons. First, by working with the "moduli surfaces" parameterizing arrangements of preimages, our problem can be reduced to the classical Diophantine problem of finding rational points on curves and surfaces. Second, our setting provides a nice example in which elliptic surfaces naturally arise and we apply specialization theorems, rank arguments, height functions, and use explicitly that the geometry of a curve has implications for its arithmetic through the use of Falting's theorem.

We make heavy use of the algebra and number theory systems Magma and PARI/gp version 2.3.2.

A similar analysis would almost certainly be possible for the families of maps of the form $x^d + c$, where $d \geq 2$ is a positive integer. In fact, for any family of polynomial maps of fixed degree it seems likely that the same methods would apply. For more general rational maps, at the very least, there would be additional complications for the genus calculations. This problem poses an interesting direction for further study.

2. Preimage curves and surfaces

In this section we summarize the necessary geometric theory of preimage curves developed in [Faber et al. 2011; 2009], and then introduce the preimage surfaces we consider in this article. Let K be a number field. As in the introduction, we define a morphism $f_c : \mathbb{A}^1_K \to \mathbb{A}^1_K$ for any $c \in K$ by the formula $f_c(x) = x^2 + c$. We could view f_c as an endomorphism of $\mathbb{P}^1_K$, but the point at infinity is totally invariant for this type of morphism and, thus, dynamically uninteresting. Fix a point $a \in K$ and a positive integer N. Define an algebraic set

$$Y^{\mathrm{pre}}(N, a) = V(f_c^N(x) - a) \subset \mathbb{A}^2_K = \operatorname{Spec} K[x, c].$$

If $Y^{\mathrm{pre}}(N, a)$ is geometrically irreducible, we define the *N-th preimage curve*, denoted $X^{\mathrm{pre}}(N, a)$, to be the unique complete curve birational to $Y^{\mathrm{pre}}(N, a)$.

Definition 2.1. We say a is an *N-th critical value* of f_c if

$$f_{c_0}^N(0) = a \quad \text{and} \quad \frac{df_c^N(0)}{dc}\bigg|_{c=c_0} = 0.$$

Theorem 2.2 [Faber et al. 2009, Corollary 2.4 and Theorem 3.2]. *Suppose N is a positive integer and $a \in K$ is not a critical value of f_c^j for any $2 \le j \le N$. Then $Y^{\mathrm{pre}}(N, a)$ is nonsingular, geometrically irreducible, and the genus of $X^{\mathrm{pre}}(N, a)$ is $(N - 3)2^{N-2} + 1$.*

For $a \in K$, define a morphism $\psi : Y^{\mathrm{pre}}(N, a) \to \mathbb{A}^N$ by

$$\psi(x, c) = \left(x, f_c(x), f_c^2(x), f_c^3(x), \dots, f_c^{N-1}(x)\right).$$

We recall the following theorem.

Theorem 2.3 [Faber et al. 2011, Proposition 4.2].

(a) *The projective closure of the image of ψ is a complete intersection of quadrics with homogenous ideal*

$$J = (Z_{N-1}^2 + Z_i Z_N - Z_{i-1}^2 - aZ_N^2 : i = 1, 2, 3, \dots, N - 1).$$

(b) *The points of $V(J)$ on the hyperplane $Z_N = 0$ have homogeneous coordinates*

$$(\epsilon_0 : \cdots : \epsilon_{N-1} : 0), \qquad \epsilon_i = \pm 1.$$

In particular, there are 2^{N-1} of them. Moreover, they are all nonsingular points of $V(J)$.

(c) *If $Y^{\mathrm{pre}}(N, a)$ is nonsingular, then $X^{\mathrm{pre}}(N, a) \cong V(J)$ and the complement of the affine part $X^{\mathrm{pre}}(N, a) \smallsetminus Y^{\mathrm{pre}}(N, a)$ consists of 2^{N-1} points.*

Definition 2.4. We define the *N-th preimage surface* $X^{\mathrm{pre}}(N)$ as the surface fibered over $\mathbb{P}^1_K$ by a. The fiber over a is given by $X^{\mathrm{pre}}(N, a)$ if $Y^{\mathrm{pre}}(N, a)$ is geometrically irreducible and $V(J)$ otherwise. In particular, for each $a \in K$ not a critical value of f_c, we get a nonsingular curve in $\mathbb{P}^N_K$.

$$\begin{tikzcd}
X^{\mathrm{pre}}(N) \arrow[d, "\pi"] & X^{\mathrm{pre}}(N, a) \arrow[d, mapsto, "\pi"] \\
\mathbb{P}^1_K & a
\end{tikzcd}$$

Note that for a fixed a_0, the affine points $(x_0, c_0, 1)$ on the curve $X^{\mathrm{pre}}(N, a_0)$ are in bijection with the N-th preimages $x_0 \in f_{c_0}^{-N}(a_0)$.

We will consider the N-th preimage surfaces in the language of function fields. In particular, consider the function field $K(a)$ which is comprised of all rational functions in a with K-rational coefficients. We consider the surfaces defined as

$$Y^{\mathrm{pre}}(N) = V(f_c^N(x) - a) \subset \mathbb{A}^2_{K(a)}$$

and

$$X^{\mathrm{pre}}(N) = V(Z_{N-1}^2 + Z_i Z_N - Z_{i-1}^2 - a Z_N^2 : i = 1, 2, 3, \ldots, N-1) \subset \mathbb{P}^N_{K(a)}.$$

The genus formula (Theorem 2.2) applies to each fiber for which $Y^{\mathrm{pre}}(N, a)$ is nonsingular and geometrically irreducible. In particular, $X^{\mathrm{pre}}(1)$ and $X^{\mathrm{pre}}(2)$ have fibers of genus 0, $X^{\mathrm{pre}}(3)$ has fibers of genus 1, and $X^{\mathrm{pre}}(N)$ for $N \geq 4$ has fibers of genus > 1 (with finitely many exceptional fibers for each N). Therefore, for $N > 3$ and all but finitely many $a \in K$, it follows from Falting's theorem that there are only finitely many points $(x, c) \in X^{\mathrm{pre}}(N, a)$. Thus, except for the finitely many a values, the N-th preimages for $N > 3$ have no contribution to $\bar{\kappa}(a)$. This premise is the content of Corollary 4.2 and the rest of Section 4 addresses the exceptional a values.

Throughout this article we discuss arrangements of preimages. For example, by a 222 arrangement we mean that there are two rational first preimages, (at least) two rational second preimages, and (at least) two rational third preimages. Similarly, a 2424 arrangement has two rational first preimages, four rational second preimages, (at least) 2 rational third preimages, and (at least) four rational fourth preimages. Note that any 226 arrangement would have to be part of a 246 arrangement since the forward image of a rational point is still a rational point.

3. Arrangements of six preimages

By examining the arrangements of six preimages we are able to prove the following lower bound for $\bar{\kappa}(a)$.

Theorem 3.1. *Let K be a number field. There is a finite set S such that*

$$\begin{cases} \bar{\kappa}(a) \geq 6 & \textit{if } a \in K \backslash (S \cap K), \\ \bar{\kappa}(a) = 4 & \textit{if } a \in S \cap K. \end{cases}$$

Proof. The 22 curve over the function field $K(a)$ is the curve whose points correspond to two rational first preimages and (at least) two rational second preimages. It has fibers of genus 0 [Faber et al. 2009] and at least one $\mathbb{Q}$-rational section for each choice of a, $(1, 1, 0)$. Thus, each fiber has infinitely many rational points and $\bar{\kappa}(a) \geq 4$.

Theorem 3.3 shows that the 222 surface has generic rank at least 2 (exactly 2 over $\mathbb{Q}$). Theorem 3.2 shows that the 24 surface has generic rank 0 over $\mathbb{Q}$. Let S be the (possibly empty) set of a values for which both the 222 and 24 surface specialize to rank 0. By [Masser and Zannier 2012] the set of a values where the 222 surface has rank 0 is finite and thus, S is finite. If $a \in S \cap K$, $\bar{\kappa}(a) = 4$, otherwise $\bar{\kappa}(a) \geq 6$. □

Second preimages. We consider the situation where the preimage tree is full to the second level; that is, there are two rational first preimages and four rational second preimages:

$$\begin{tikzcd}
 & & & a & & & . \\
 & t \arrow[urr, "f_c"] & & & & -t \arrow[ull, "f_c"'] & \\
s \arrow[ur, "f_c"] & & -s \arrow[ul, "f_c"'] & & u \arrow[ur, "f_c"] & & -u. \arrow[ul, "f_c"']
\end{tikzcd}$$

We can define this curve over the function field $K(a)$ as

$$X_{24} = V(s^2 - tz - (t^2 - az^2), u^2 + tz - (t^2 - az^2)) \subseteq \mathbb{P}^3_{K(a)}.$$

The fibers (when nonsingular) have genus one with at least one rational section $(1, 1, 1, 0)$ so we can produce a minimal Weierstrass model (using Magma) as an elliptic curve over the function field $K(a)$ as

$$E_{24}(a) : v^2 w = u^3 + (4a - 1)u^2 w + 16auw^2 + (64a^2 - 16a)w^3$$

with j-invariant

$$j(a) = \frac{(16a^2 - 56a + 1)^3}{a(4a + 1)^4}$$

and discriminant

$$\Delta(a) = a(4a + 1)^4.$$

The only fibers which are not elliptic curves are $a = 0$ and $a = -\frac{1}{4}$. This is in fact a rational elliptic surface since it has a Weierstrass model satisfying $\deg(a_i) \leq i$ for a_i the coefficients of an elliptic curve in Weierstrass form [Shioda 1990, page 237].

Theorem 3.2. *$E_{24}(a)(\mathbb{Q}(a))$ has rank 0 and torsion subgroup $\mathbb{Z}/4\mathbb{Z}$ generated by*

$$T(a) = (2, 8a + 2, 1).$$

Proof. We use the main theorem of [Oguiso and Shioda 1991] to see that the rank over $\mathbb{Q}(a)$ is zero. We compute the Kodaira symbols in Magma to get

$$[\langle I4, 1\rangle\, , \langle I1{*}, 1\rangle\, , \langle I1, 1\rangle\,].$$

From row 72 in the table [Oguiso and Shioda 1991] we have that the rank of $E_{24}(a)(\mathbb{Q}(a))$ is zero. Examining the torsion, we see that the point

$$(2, 8a+2, 1)$$

has order 4 and the specialization $E_{24}(1)(\mathbb{Q})$ has torsion subgroup $\mathbb{Z}/4\mathbb{Z}$. Since the specialization map is injective on torsion on all nonsingular fibers, $E_{24}(a)$ has torsion subgroup exactly $\mathbb{Z}/4\mathbb{Z}$. □

Third preimages. From Theorem 2.3 we see that the elliptic surface parameterizing third preimages of a over the function field $K(a)$ is given by

$$X_{222} = V(z_2^2 + z_1z_3 - z_0^2 - az_3^2, z_2^2 + z_2z_3 - z_1^2 - az_3^2) \subseteq \mathbb{P}^3_{K(a)}.$$

Using the cuspidal point $(-1, 1, 1, 0)$ from Theorem 2.3 as the section at infinity we can find a minimal model in Magma as

$$E_{222}(a) : v^2w = u^3 + \left(16a + \tfrac{942}{13}\right)u^2w + \left(\tfrac{10048}{13}a + \tfrac{293084}{169}\right)uw^2 + \left(1024a^2 + \tfrac{1620800}{169}a + \tfrac{30250696}{2197}\right)w^3$$

with j-invariant

$$j(a) = \frac{(16a^2+3)^2}{(4a+1)^2(256a^3+368a^2+104a+23)}$$

and discriminant

$$\Delta(a) = (4a+1)^2(256a^3+368a^2+104a+23).$$

As expected, the only fibers which are not elliptic curves are the fibers over $a = -\frac{1}{4}$ and the three third critical values. This is in fact a rational elliptic surface since it has a Weierstrass model satisfying $\deg(a_i) \le i$ for a_i the coefficients of an elliptic curve in Weierstrass form [Shioda 1990, page 237].

Theorem 3.3. *$E_{222}(a)(\mathbb{Q}(a))$ has rank 2 generated by the two independent sections*

$$P(a) = \left(-\tfrac{262}{13}, 32a+8, 1\right) \quad \textit{and} \quad Q(a) = \left(-\tfrac{366}{13}, 32a+8, 1\right).$$

Proof. We use the main theorem of [Oguiso and Shioda 1991] to see that the rank over $\mathbb{Q}(a)$ is exactly two. We compute the Kodaira symbols in Magma to get

$$[\langle I1, 3\rangle\, , \langle I2, 1\rangle\, , \langle I1{*}, 1\rangle\,].$$

From row 30 in the table [Oguiso and Shioda 1991] we have that the rank of $E_{222}(a)(\mathbb{Q}(a)) = 2$. Since the specialization map is injective on torsion on all fibers where E_{222} is nonsingular, and the specialization $E_{222}(0)$ has no torsion, there are no rational torsion sections. We can see $P(a)$ and $Q(a)$ are actually the generators by finding a specialization $E_{222}(a_0)$ which is rank 2 with generators $P(a_0)$ and $Q(a_0)$. For $a = 4$ we have

$$E_{222}(4) : v^2 w = u^3 + \tfrac{1774}{13} u^2 w + \tfrac{815580}{169} u w^2 + \tfrac{150527944}{2197} w^3$$

and from Magma the generators are

$$\left(-\tfrac{262}{13}, 136, 1\right) \quad \text{and} \quad \left(-\tfrac{1146}{13}, 136, 1\right).$$

In terms of $P(4)$ and $Q(4)$ these are

$$P(4) \quad \text{and} \quad P(4) + Q(4).$$

Thus, $P(4)$ and $Q(4)$ generate the Mordell-Weil group $E_{222}(4)$ and, hence, $P(a)$ and $Q(a)$ generate the Mordell-Weil group of $E_{222}(a)$. □

4. Arrangements of eight or more preimages

We examine when the genus of the fibers of preimage surfaces of various arrangements of $2N$ preimages is greater than 1 and, thus, by Falting's theorem have a finite number of rational points over an algebraic number field. In particular, if every $2N$ arrangement has genus greater than 1 for some N, then $\bar{\kappa}(a) < 2N$. The difficulty lies in determining the genus when the fiber is singular. We treat the nonsingular case in the following theorem.

Theorem 4.1. *If the curve (fiber) defining an arrangement of $2N$ rational preimages of a is nonsingular, then it has genus $(N-3)2^{N-2} + 1$.*

Proof. A complete intersection in $\mathbb{P}^m$ is defined as a subscheme Y of $\mathbb{P}^m$ whose homogeneous ideal I can be generated by $r = \operatorname{codim}(Y, \mathbb{P}^m)$ elements [Hartshorne 1977, Exercise II.8.4]. Each surface arranging $2N$ points can be described by the equations

$$f_c(z_1) = a \qquad \text{and} \qquad f_c(z_i) = (-1)^{\epsilon} z_j \text{ for } 2 \le i \le N$$

where $1 \le j < N$ and $\epsilon = \pm 1$ depending on the arrangment of points. After homogenization and elimination of c from this system of equations we obtain a description of each fiber as a curve defined by $N-1$ degree two hypersurfaces in $\mathbb{P}^N$ and, hence, a complete intersection. From [Hirzebruch 1966, §22] or [Arslan and Sertöz 1998, Corollary 2] we get a formula for the arithmetic genus of a complete intersection of $N-1$ degree two hypersurfaces in $\mathbb{P}^N$ as

$$p_a = \sum_{m=1}^{N-1} (-1)^{m+1} \binom{N-1}{m} \phi_N(-2m)$$

where $\phi_N(z)$ comes from the Hilbert polynomial of the $2N$ curve and is given by

$$\phi_N(z) = \frac{(z+1)(z+2)\cdots(z+N)}{N!} = \binom{z+N}{N}.$$

Since the arithmetic genus is equal to the geometric genus for nonsingular curves [Hartshorne 1977, Proposition IV.1.1], the genus is independent of the arrangement of the preimages and from [Faber et al. 2009, Theorem 1.5] we get the simpler formula

$$g = (N-3)2^{N-2} + 1. \qquad \square$$

Corollary 4.2. *If the curve (fiber) defining an arrangement of* $2N$ *rational preimages of a is nonsingular, then the genus is greater than* 1 *for* $2N \geq 8$.

We have thus reduced the computation of $\bar{\kappa}(a, K)$ to checking a values where the fiber is singular for arrangements with 8 (or more) rational preimages (224, 242, 2222). The method is as follows.

(a) Using the Jacobian criterion, determine all of the singular fibers (a values).

(b) Determine the δ-invariants of each singular point to determine the genus of each singular fiber.

Recall that the δ-invariant of a singularity P is defined as

$$\delta_P = \sum_{Q} \tfrac{1}{2} m_Q(m_Q - 1),$$

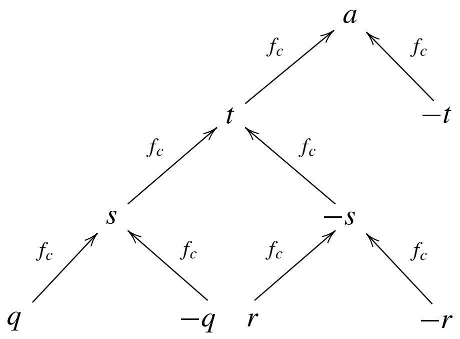
where the sum ranges over the infinitely near points of P and m_Q are their multiplicities. See [Sendra et al. 2008, Section 3.2] for the basic definitions and the case of plane curves and [Brieskorn and Knörrer 1986, Section 9.2, Theorem 7] for a more general discussion. As the singularity analysis computations are identical in form for all of the singularities, we outline the method, include the first such computation, and omit the details for the other singularities. The singularity analysis proceeds as follows.

(a) Let $C \subseteq \mathbb{P}^N$ be a singular curve with singular point P. We move P to $(0, \dots, 0, 1)$ and dehomogenize.

(b) Project onto a singular plane curve with isomorphic tangent space at the singular point.

(c) Analyze the singularity of the plane curve with blow-ups and compute the δ-invariant.

Examining the 224 surface. One possible 224 arrangement of 8 preimages is this:



Every other 224 arrangement differs only by renaming, so this is the only distinct 224 arrangement. The curve is defined by three degree two equations in $\mathbb{P}^4$ as

$$C_{224} = V(az^2 - t^2 - (tz - s^2), az^2 - t^2 - (sz - q^2), az^2 - t^2 - (-sz - r^2)) \subseteq \mathbb{P}^4_{K(a)}.$$

Theorem 4.3. *The a values for which the fiber of the* 224 *surface is singular are given by*

$$a \in \left\{-\tfrac{1}{4}, 0, a_1, a_2, a_3\right\},$$

where a_1, a_2, a_3 are the three third critical values of f_c

Proof. We apply the Jacobian criterion to determine the singular points. For each singular point, we can determine the associated a value(s). Examining the hyperplane at infinity $z = 0$ we have the 8 cuspidal points $(\pm 1, \pm 1, \pm 1, 1, 0) \in \mathbb{P}^4$. To check the singularity of these points, we use the Jacobian criterion on the affine chart $\mathbb{A}^4_{q \neq 0}$ with generators

$$\{az^2 - t^2 - (tz - s^2), az^2 - t^2 - (sz - 1), az^2 - t^2 - (-sz - r^2)\}$$

to have the Jacobian matrix at $z = 0$

$$\begin{pmatrix} 0 & 2s & -2t & -t \\ 0 & 0 & -2t & -s \\ 2r & 0 & -2t & s \end{pmatrix}.$$

The determinant of one such maximal minor is $-8rst$, and since $r, s, t \neq 0$, this is nonzero, so the cuspidal points are all nonsingular.

Now we consider the points in the affine chart $\mathbb{A}^4_{z \neq 0}$ which has generators

$$\{a - t^2 - (t - s^2), a - t^2 - (s - q^2), a - t^2 - (-s - r^2)\}.$$

The Jacobian matrix is given by

$$\begin{pmatrix} 0 & 0 & 2s & -2t-1 \\ 2q & 0 & -1 & -2t \\ 0 & 2r & 1 & -2t \end{pmatrix}$$

and the determinants of the maximal minors are

$$\{8qrs, 4qr(-2t-1), 2q(4st-2t-1), -2r(4st+2t+1)\}.$$

The combinations that result in all 4 determinants vanishing are the following.

(a) If $q = r = 0$, then we have $c = \pm s$ and so $c = 0$ and so $a = 0$.

(b) If $q = 0$ and $(4st+2t+1) = 0$, then we must have $s \neq -\frac{1}{2}$ so we can solve $t = -\frac{1}{4s+2} = -\frac{1}{4c+2}$. Then we have $s^2 + c = c^2 + c = t$ and the roots of $4c^3 + 6c^2 + 2c + 1 = \frac{df_c^3(0)}{dc}$ combined with $a = f_c(f_c(f_c(0)))$ to get the three third critical values.

(c) If $q \neq 0$, $r = 0$, and $(4st - 2t - 1) = 0$, then we must have $t \neq 0$ and we can solve $s = \frac{2t+1}{4t} = -c$. Then we have $s^2 - s = t$ and the roots of $16t^3 + 4t^2 - 1$ which give the three third critical values.

(d) If $q, r \neq 0$, $s = 0$, and $t = -\frac{1}{2}$, then we have $c = -\frac{1}{2}$ and so $a = -\frac{1}{4}$.

□

We will treat $a = -\frac{1}{4}$ on page 358.

Theorem 4.4. *The genus of* C_{224} *is*

$$g = \begin{cases} 4 & \textit{if } a = 0, \\ 1 & \textit{if } a \in \{a_1, a_2, a_3\}, \end{cases}$$

where a_1, a_2, a_3 *are the three third critical values of* f_c.

Proof. There is one singular point for $a = 0$ and four singular points for each a_i. In all cases $\delta_P = 1$ so the genus drops by 1 for each singular point.

We now compute the δ-invariant of one of the singular points for a_1. The 224 curve for a_1 is defined as

$$V(a_1z^2 - t^2 - (tz - s^2), a_1z^2 - t^2 - (sz - q^2), a_1z^2 - t^2 - (-sz - r^2))$$

and if α is a root of

$$4x^3 + 6x^2 + 2x + 1$$

then

$$a_1 = \alpha^4 + 2\alpha^3 + \alpha^2 + \alpha = -\tfrac{1}{4}\alpha^2 + \tfrac{1}{2}\alpha - \tfrac{1}{8}.$$

We label the coordinates as (q, r, s, t, z) and the singular point is

$$P = (0, -\beta, \alpha, \alpha^2 + \alpha, 1)$$

where $\beta^2 = -2\alpha$. We move P to $(0, 0, 0, 0, 1)$ with a translation

$$(q, r, s, t, z) \mapsto (q, r - \beta z, s + \alpha z, t + (\alpha^2 + \alpha)z)$$

to get a new curve $\widetilde{C}$ and singular point $\widetilde{P} = (0, 0, 0, 0, 1)$. We dehomogenize to affine coordinates $(Q, R, S, T) = (q/z, r/z, s/z, t/z)$ and compute the tangent space at $\widetilde{P}$ as

$$\begin{cases} -2T\alpha^2 - 2T\alpha - T + 2S\alpha = 0, \\ -2T\alpha^2 - 2T\alpha - S = 0, \\ -2T\alpha^2 - 2T\alpha + S - 2\beta R = 0. \end{cases} \tag{1}$$

Notice that the second equation of (1) implies the first using the degree 4 polynomial satisfied by α. Thus, the tangent space is given by

$$-2T\alpha^2 - 2T\alpha - S = 0, \quad -2T\alpha^2 - 2T\alpha + S - 2\beta R = 0.$$

Since we want to project $\widetilde{C}$ to a plane curve preserving the tangent space at $\widetilde{P}$ we define

$$u = -2T\alpha^2 - 2T\alpha - S, \quad v = -2T\alpha^2 - 2T\alpha + S - 2\beta R,$$

with inverse

$$S = \beta R - \frac{u}{2} + \frac{v}{2}, \quad T = \frac{\beta R}{-2\alpha^2 - 2\alpha} + \frac{u}{-4\alpha^2 - 4\alpha} + \frac{v}{-4\alpha^2 - 4\alpha},$$

and make the change of variables $(Q, R, S, T) \mapsto (Q, R, u, v)$ to get a new curve $\widetilde{C}'$ and point $\widetilde{P}'$. The tangent space at $\widetilde{P}'$ is given by $u = v = 0$. We now project $\widetilde{C}'$ onto a plane curve in the QR-plane. To project we eliminate the variables u, v from the three defining equations of $\widetilde{C}'$ to get the single equation

$$\begin{aligned}
&(2\alpha + 1)Q^8 + \big((-8\alpha - 4)R^2 + (16\beta\alpha + 8\beta)R + (16\alpha^2 - 4)\big)Q^6 \\
&\quad + \big((12\alpha + 6)R^4 + (-48\beta\alpha - 24\beta)R^3 + (-144\alpha^2 - 64\alpha + 4)R^2 \\
&\qquad\qquad + (96\beta\alpha^2 + 32\beta\alpha - 8\beta)R + (-64\alpha^2 - 24\alpha - 8)\big)Q^4 \\
&\quad + \big((-8\alpha - 4)R^6 + (48\beta\alpha + 24\beta)R^5 + (240\alpha^2 + 128\alpha + 4)R^4 \\
&\qquad + (-320\beta\alpha^2 - 192\beta\alpha - 16\beta)R^3 + (384\alpha^2 + 208\alpha + 128)R^2 \\
&\qquad\qquad + (-128\beta\alpha^2 - 96\beta\alpha - 64\beta)R - 32\alpha\big)Q^2 \\
&\quad + (2\alpha + 1)R^8 + (-16\beta\alpha - 8\beta)R^7 + (-112\alpha^2 - 64\alpha - 4)R^6 \\
&\qquad + (224\beta\alpha^2 + 160\beta\alpha + 24\beta)R^5 + (-320\alpha^2 - 152\alpha - 136)R^4 \\
&\qquad\qquad + (128\beta\alpha^2 + 32\beta\alpha + 96\beta)R^3 + (-64\alpha^2 + 64\alpha)R^2 = 0,
\end{aligned}$$

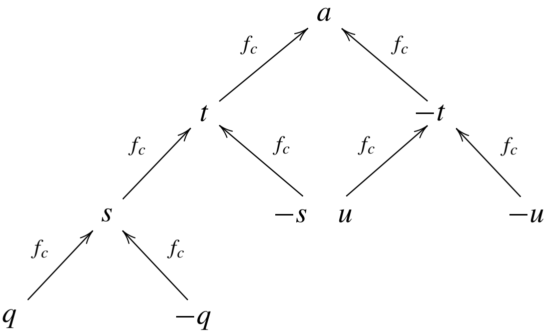
defining a plane curve in $\mathbb{A}^2$ with variables (Q, R). Notice that the only points of the form $(0, 0, u, v)$ on $\widetilde{C}'$ is the point $(0, 0, 0, 0)$ (in the other words, the singular point is the only point that projects onto $(0, 0)$), so we proceed with analyzing the plane curve singularity $(0, 0)$. Blowing-up once resolves the singularity and we see that it has multiplicity 2. So we compute

$$\delta_P = \frac{1}{2}(2 \cdot 1) = 1.$$

A similar analysis is done on all of the other singularities to get $\delta_P = 1$ for all P for all $a \in \{0, a_1, a_2, a_3\}$. Hence, we have

$$\begin{cases} g = 5 - 1 = 4 & \text{if } a = 0, \\ g = 5 - (1 + 1 + 1 + 1) = 1 & \text{if } a = a_1, a_2, a_3. \end{cases} \qquad \square$$

Examining the 242 surface. One possible 242 arrangement of 8 preimages is this:



Every other 242 arrangement differs only by renaming, so this is the only distinct 242 arrangement. The surface is defined by 3 degree two equations in $\mathbb{P}^4$ as

$$C_{242} = V\big(az^2 - t^2 - (tz - s^2), az^2 - t^2 - (-tz - u^2), az^2 - t^2 - (sz - q^2)\big) \subseteq \mathbb{P}^4_{K(a)}.$$

Theorem 4.5. *The a values for which the fiber of the* 242 *surface is singular are given by*

$$a \in \left\{-\tfrac{1}{4}, 0, 2, a_1, a_2, a_3\right\}$$

where a_1, a_2, a_3 are the three third critical values of f_c.

Proof. We apply the Jacobian criterion to determine the singular points. For each singular point, we can determine the associated a value(s). Examining the hyperplane at infinity, $z = 0$, we have the 8 cuspidal points $(\pm 1, \pm 1, \pm 1, 1, 0) \in \mathbb{P}^4$. To check the singularity of these points, we use the Jacobian criterion on the affine chart $\mathbb{A}^4_{q \neq 0}$ with generators

$$\big\{az^2 - t^2 - (tz - s^2), az^2 - t^2 - (-tz - u^2), az^2 - t^2 - (sz - 1)\big\}.$$

The Jacobian matrix at $z = 0$ is given by

$$\begin{pmatrix} 2s & 0 & -2t & -t \\ 0 & 2u & -2t & t \\ 0 & 0 & -2t & -s \end{pmatrix}.$$

The determinant of one maximal minor is $-8sut$, and since $s, u, t \neq 0$, this is nonzero, so the cuspidal points are all nonsingular.

Now we consider the points in the affine chart $\mathbb{A}^4_{z\neq 0}$ which has generators

$$\{a - t^2 - (t - s^2), a - t^2 - (-t - u^2), a - t^2 - (s - q^2)\}.$$

The Jacobian matrix is given by

$$\begin{pmatrix} 0 & 2s & -2t-1 & 0 \\ 0 & 0 & -2t+1 & 2u \\ 2q & -1 & -2t & 0 \end{pmatrix}.$$

The determinants of the maximal minors are

$$\{2u(4st + 2t + 1), 4qu(-2t - 1), 8qus, 4qs(-2t + 1)\}.$$

The combinations that result in all 4 vanishing are as follows:

(a) If $q = 0$ and $u = 0$, then $f_c^2(0) = a$ and $f_c^3(0) = a$ which is the polynomial equation

$$f_c(f_c(f_c(0))) - f_c(f_c(0)) = c^4 + 2c^3 = c^3(c + 2) = 0$$

so $c = 0$ or $c = -2$. So we have $a = 0$ or $a = 2$.

(b) If $q = 0$ and $(4st + 2t + 1) = 0$, then we must have $s \neq -\frac{1}{2}$ so we can solve $t = -1/(4s + 2) = -1/(4c + 2)$. Then we have $s^2 + c = c^2 + c = t$ and the roots of

$$4c^3 + 6c^2 + 2c + 1 = \frac{df_c^3(0)}{dc}$$

combined with $a = f_c(f_c(f_c(0)))$ to get the three third critical values.

(c) If $u = 0$ and $s = 0$, then $c = \pm t$ and so $t = c = 0$ and so $a = 0$.

(d) If $u = 0$ and $t = \frac{1}{2}$, then $c = -\frac{1}{2}$ and so $a = -\frac{1}{4}$.

(e) If $s = 0$ and $t = -\frac{1}{2}$, then $c = -\frac{1}{2}$ and so $a = -\frac{1}{4}$. □

We will treat $a = -\frac{1}{4}$ on page 358.

Theorem 4.6. *The genus of* C_{242} *is* $g = \begin{cases} 3 & \textit{if } a = 0, \\ 4 & \textit{if } a = 2, \\ 3 & \textit{if } a \in \{a_1, a_2, a_3\}. \end{cases}$

Proof. We proceed as in the proof of Theorem 4.4 for analyzing the singularities.

For $a = 0$ there is one singularity that required two blow-ups to resolve and we get multiplicity 2 for both of the infinitely near points and, hence, $\delta_P = \frac{1}{2}(2 \cdot 1) + \frac{1}{2}(2 \cdot 1) = 2$ and $g = 5 - 2 = 3$.

For $a = 2$ there is one singular point with $\delta_P = 1$ and, hence, $g = 5 - 1 = 4$.

For $a \in \{a_1, a_2, a_3\}$ each curve has two singular points both with $\delta_P = 1$ and, hence, $g = 5 - (1 + 1) = 3$. □

Examining the 2222 surface. One possible 2222 arrangement of 8 preimages is this:

$$\begin{tikzcd}
 & & & & a & \\
 & & & t \arrow[ur, "f_c"] & & -t \arrow[ul, "f_c"'] \\
 & & s \arrow[ur, "f_c"] & & -s \arrow[ul, "f_c"'] & \\
 & q \arrow[ur, "f_c"] & & -q \arrow[ul, "f_c"'] & & \\
u \arrow[ur, "f_c"] & & -u \arrow[ul, "f_c"'] & & &
\end{tikzcd}$$

Every other 2222 arrangement differs only by renaming, so this is the only distinct 2222 arrangement. The surface is defined by 4 degree two equations in $\mathbb{P}^5$ as

$$C_{2222} = V(az^2 - t^2 - (tz - s^2), az^2 - t^2 - (sz - q^2), az^2 - t^2 - (qz - u^2)) \subseteq \mathbb{P}^5_{K(a)}.$$

From [Faber et al. 2009, Theorem 1.3] the only singular fibers are for a the N-th critical values for $2 \le N \le 4$. For $N = 2$ we get $a = -\frac{1}{4}$, which will be treated on page 358. For $N = 3$ we get the three third critical values which we label $a_{3,1}, a_{3,2}, a_{3,3}$. For $N = 4$ we get the seven 4-th critical values, which we label $a_{4,i}$ for $1 \le i \le 7$, and which satisfy

$$a = f_c(f_c(f_c(f_c(0)))) \quad \text{for } 8c^7 + 28c^6 + 36c^5 + 30c^4 + 20c^3 + 6c^2 + 2c + 1 = 0.$$

Theorem 4.7. *The genus of C_{2222} is*

$$g = \begin{cases} 3 & \text{if } a \in \{a_{3,1}, a_{3,2}, a_{3,3}\}, \\ 4 & \text{if } a \in \{a_{4,i} : 1 \le i \le 7\}. \end{cases}$$

Proof. A fiber of the 2222 surface is isomorphic [Faber et al. 2011, Proposition 4.2] to the degree 16 plain curve defined by the equation

$$f_c^4(x) = a.$$

For $a \in \{a_{3,i}\}$ there are three singular points, one of which is $(0, 1, 0)$ and the other two depend on a. The $(0, 1, 0)$ point requires several blow-ups and has $\delta_P = 100$ and each of the other two points have $\delta_P = 1$ for a final genus of $g = \frac{1}{2}(15 \cdot 14) - 102 = 105 - 102 = 3$.

For $a \in \{a_{4,i}\}$ there are two singular points, one of which is $(0, 1, 0)$ and the other depends on a. The $(0, 1, 0)$ point has $\delta_P = 100$ and the point has $\delta_P = 1$ for a final genus of $g = \frac{1}{2}(15 \cdot 14) - 101 = 105 - 101 = 4$. □

Corollary 4.8. *For any $a \in \overline{\mathbb{Q}} \backslash \{-\frac{1}{4}\}$ and any algebraic number field K there are only finitely many $c \in K$ for which there are at least two K-rational 4-th preimages of a.*

The bound $\bar{\kappa}(-\frac{1}{4})$. For $a = -\frac{1}{4}$ the preimages curves are in fact reducible since we have an equation in the generators of the form

$$s^2 + \left(t - \tfrac{1}{2}z\right)^2 = \left(s - \left(t - \tfrac{1}{2}z\right)\right)\left(s + \left(t - \tfrac{1}{2}z\right)\right),$$

where s is a second preimage of a for which $s^2 + c = t$ and $t^2 + c = a$, and an equation of the form

$$u^2 - \left(t + \tfrac{1}{2}z\right)^2 = \left(u - \left(t + \tfrac{1}{2}z\right)\left(u - \left(t + \tfrac{1}{2}z\right)\right),\right.$$

where u is a second preimage of a for which $u^2 + c = -t$. After splitting the preimage curves into their distinct irreducible components we can again proceed with genus calculations.

Theorem 4.9. *For any fixed number field K, $\bar{\kappa}\left(-\frac{1}{4}\right) = 10$.*

Proof. Using the Jacobian criterion we compute that the following curves are all nonsingular, and we apply the genus formula from [Hirzebruch 1966, §22] or [Arslan and Sertöz 1998, Corollary 2] to compute the following genera.

$$g = \begin{cases} 1 & \text{in the cases } 224, 2222, 244, 2422 \\ 5 & \text{in the cases } 22222, 2224, 2242, 246, 2442, 2424, 24222. \end{cases}$$

Using Magma, we see that the 244 curve is a rank 1 elliptic curve over $\mathbb{Q}$ isomorphic to

$$v^2 w = u^3 + u^2 w - 9uw^2 + 7w^3$$

so has infinitely many rational points. Therefore, there are infinitely many c with 10 rational preimages of $-\frac{1}{4}$ and only finitely many c values with 12 (or more) rational preimages of $-\frac{1}{4}$. □

5. Proof of Theorem 1.3

Proof. The case $a = -\frac{1}{4}$ was covered in Theorem 4.9.

For a a third critical value we have genus 1 for the 224 curve and, hence, for a large enough extension of $\mathbb{Q}$ it has positive rank and infinitely many rational points. Also, it has no $\mathbb{Q}$-rational points. The 242 curve has genus greater than 1 and, hence, has only finitely many rational points. Thus, for $\bar{\kappa}(a, K)$ to be at least 10 there must be infinitely many rational points on a curve corresponding to an arrangement with rational 4-th preimages, which is not possible by Corollary 4.8. So it is possible for $\bar{\kappa}(a, K)$ to be either 6 or 8 depending on the field.

For all other values of a we have the genus of the 224 and 242 curves are greater than 1 and, hence, have only finitely many rational points. Any arrangement with more points must contain one of these two arrangements, hence $\bar{\kappa}(a, K) \leq 6$. Theorem 3.3 shows that the 222 surface has generic rank 2 and [Masser and Zannier 2012] shows that the set of a where the rank is 0 is finite. Every a value for which both E_{222} and E_{24} specialize to rank 0 has $\bar{\kappa}(a) = 4$, otherwise $\bar{\kappa}(a) = 6$. □

6. Other properties of preimage surfaces

In this section we collect some additional properties of the preimages surfaces that are tangential to the proof of Theorem 1.3, yet still of interest.

Parametrization of torsion subgroups of E_{24}. Recall that Mazur's theorem [1977] gives a description of the possible torsion subgroups of elliptic curves over $\mathbb{Q}$ and that the specialization map is injective on nonsingular fibers. These facts combined with Theorem 3.2 implies that the possible torsion subgroups for a nonsingular specialization of $E_{24}(a)$ must be isomorphic to one of the following groups:

$$\{\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z},\ \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/8\mathbb{Z},\ \mathbb{Z}/4\mathbb{Z},\ \mathbb{Z}/8\mathbb{Z},\ \mathbb{Z}/12\mathbb{Z}\}.$$

We characterize the a values giving rise to a specialization with each of these possible torsion subgroups in the following theorem.

Theorem 6.1. (a) *$E_{24}(a)(\mathbb{Q})$ contains a subgroup isomorphic to $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z}$ if and only if*

$$a = -t^2 \quad \textit{for} \quad t \in \mathbb{Q}\backslash\{0, \pm\tfrac{1}{2}\}.$$

(b) *$E_{24}(a)(\mathbb{Q})$ contains a subgroup isomorphic to $\mathbb{Z}/8\mathbb{Z}$ if and only if*

$$a = \tfrac{1}{4}t^2(t^2 - 2) \quad \textit{for} \quad t \in \mathbb{Q}\backslash\{0, \pm 1\}.$$

(c) *$E_{24}(a)(\mathbb{Q})$ contains a subgroup isomorphic to $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/8\mathbb{Z}$ if and only if*

$$a = -\frac{(4t^2 - 4t - 1)^2(4t^2 + 4t - 1)^2}{4(4t^2 + 1)^4} \quad \textit{for} \quad t \in \mathbb{Q}\backslash\{0, \pm\tfrac{1}{2}\}.$$

(d) *$E_{24}(a)(\mathbb{Q})$ contains a subgroup isomorphic to $\mathbb{Z}/12\mathbb{Z}$ if and only if*

$$a = \frac{(13691470144t^2 - 235376t + 1)(13903463744t^2 - 235376t + 1)^3}{9527265101250297856000000t^6(117688t - 1)^2}$$

for $t \in \mathbb{Q}\backslash\left\{0, \frac{1}{117688}\right\}$.

Proof. (a) First suppose $a = -t^2$ for some $t \in \mathbb{Q}\backslash\left\{0, \pm\frac{1}{2}\right\}$. Then

$$\{\mathbb{O}, (4t^2+1, 0, 1), (4t, 0, 1), (-4t, 0, 1)\}$$

is a subgroup of $E_{24}(-t^2)(\mathbb{Q})$ isomorphic to $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$. Since there is also a generic torsion point of order 4 (Theorem 3.2), $E_{24}(-t^2)(\mathbb{Q})$ contains a subgroup isomorphic to $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z}$. Next, suppose $E_{24}(a)(\mathbb{Q})$ contains a subgroup isomorphic to $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ and, hence, also a subgroup isomorphic to $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z}$. Thus, $E_{24}(a)(\mathbb{Q})$ has three points of order two. Points of order two must be rational roots of the Weierstrass equation

$$x^3 + (4a-1)x^2 + (16a)x + 16a(4a-1) = (x+4a-1)(x^2+16a). \tag{2}$$

So, x^2+16a must have 2 rational roots, or equivalently, $a = -(x/4)^2 = -t^2$. Hence, there are three rational roots of (2) if and only if $a = -t^2$ for $t \in \mathbb{Q}$. However, if $t = \pm\frac{1}{2}$ then the roots will not be distinct, so we must have $a = -t^2$ for $t \in \mathbb{Q}\backslash\{\pm\frac{1}{2}\}$. For $t = 0$ we get $a = 0$ which is a degenerate case (a singular fiber of $X^{\text{pre}}(2)$).

(b) Suppose $a = t^2(t^2-2)/4$ for some $t \in \mathbb{Q}\backslash\{0, \pm 1\}$. Then it can be verified directly that the point $P = (2t(t^2+t-1), 2(t-1)t(t+1)^3, 1)$ is in $E_{24}(a)(\mathbb{Q})$ and $[2]P = (2, 2(4a+1), 1)$ is the generator of the cyclic subgroup of order four. So, P generates a cyclic group of order eight.

Now suppose that $E_{24}(a)(\mathbb{Q})$ has a cyclic subgroup of order eight. If we let $P = (x, y, 1)$ be the generator of the subgroup, then $[2]P$ generates a cyclic group of order four (the generic torsion subgroup). So, we must have $x([2]P) = 2$. This gives us the equation

$$x^4 - 8x^3 - 64ax^2 + 8x^2 - 512a^2x - 1024a^3 + 256a^2 + 64a = 0.$$

Then using the solution to the quartic we have the solutions

$$x = 2 \pm 2\sqrt{4a+1} + \frac{1}{2}\sqrt{24 + (8a-1) \pm \frac{512 + 4096a^2 + 256(8a-1)}{16\sqrt{4a+1}}}$$
$$x = 2 \pm 2\sqrt{4a+1} - \frac{1}{2}\sqrt{24 + (8a-1) \pm \frac{512 + 4096a^2 + 256(8a-1)}{16\sqrt{4a+1}}}.$$

In order to have $x \in \mathbb{Q}$, and since x is clearly not 2, we must have $\sqrt{4a+1} \in \mathbb{Q}$. So $a = \frac{b^2-1}{4}$ for some $b \in \mathbb{Q}$. The above roots become

$$x = 2(1 \pm b + b\sqrt{1 \pm b})$$
$$x = 2(1 \pm b - b\sqrt{1 \pm b})$$

from which it follows that $b = \pm(t^2 - 1)$. Thus, $a = \frac{t^2(t^2-2)}{4}$. Note that for $t = \pm 1$ we get $a = -\frac{1}{4}$ and for $t = 0$ we get $a = 0$ which are all singular fibers.

(c) Clearly, $E_{24}(a)(\mathbb{Q})$ has a subgroup isomorphic to $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/8\mathbb{Z}$ if and only if $E_2(a)$ has a subgroup isomorphic to $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z}$ and a subgroup isomorphic to $\mathbb{Z}/8\mathbb{Z}$. From the two previous parts, it follows that $a = -t_1^2$ and $a = \frac{1}{4}t_2^2(t_2^2 - 2)$. These two equations define a curve of genus zero which can be parameterized with Magma and substituted into $a = -t_1^2$ to get the stated form. For $t = 0, \pm\frac{1}{2}$ we get $a = -\frac{1}{4}$, which is a singular fiber.

(d) Since specialization is injective on torsion for nonsingular fibers , $E_{24}(a)(\mathbb{Q})$ has a subgroup isomorphic to $\mathbb{Z}/12\mathbb{Z}$ if and only if there is a point $Q = [x, y] \in E_{24}(a)(\mathbb{Q})$ for which $[3]Q$ generates the generic $\mathbb{Z}/4\mathbb{Z}$ torsion subgroup. In particular, we must have $x([3]Q) = 2$. So we need to find solutions to

$$\frac{x([3]Q) - 2}{x - 2} = 0$$

where we divide out by $x-2$ since we only wish to exclude the a values which have purely $\mathbb{Z}/4\mathbb{Z}$ torsion. From the *algcurve* package in Maple we get the parametrization given. The two excluded t values correspond to the two singular fibers $a = 0$ and $a = -\frac{1}{4}$. □

Corollary 6.2. *The $a \in \mathbb{Q}$ for which $E_{24}(a)(\mathbb{Q})$ has torsion subgroup exactly $\mathbb{Z}/4\mathbb{Z}$, in other words, the $a \in \mathbb{Q}$ for which the specialization map is an isomorphism on torsion, is a Zariski dense set.*

Proof. From Mazur's theorem and the injectivity of the specialization map, the possible torsion groups of $E_{24}(a)(\mathbb{Q})$ are

$$\{\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z},\ \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/8\mathbb{Z},\ \mathbb{Z}/4\mathbb{Z},\ \mathbb{Z}/8\mathbb{Z},\ \mathbb{Z}/12\mathbb{Z}\}.$$

The condition on a for $E_{24}(a)(\mathbb{Q})_{\text{tors}}$ to not be $\mathbb{Z}/4\mathbb{Z}$ is a closed condition from Theorem 6.1 and the j-invariant. Therefore, every $a \in \mathbb{Q}$ outside of this Zariski closed set satisfies $E_{24}(a)(\mathbb{Q})_{\text{tors}} \cong \mathbb{Z}/4\mathbb{Z}$ and there is at least one such a,

$$E_{23}(1)(\mathbb{Q})_{\text{tors}} \cong \mathbb{Z}/4\mathbb{Z}.$$ □

Exceptional (c, a) values over $\mathbb{Q}$.

Rank zero. The methods of [Masser and Zannier 2008; 2012], in principle, can compute the full set S, but in practice such computations are difficult. However, computing the set $S \cap K$ for $[K : \mathbb{Q}] \leq 2$ from Theorem 1.3 is feasible since we have an explicit (small) bound on the order of a torsion point.

We must have both $P(a)$ and $Q(a)$ are torsion on the 222 surface. We have a bound of 18 for the order of a torsion point over a quadratic number field K [Kamienny 1992; Kenku and Momose 1988]. Finding the a for which $P(a)$ or $Q(a)$ is torsion of a given order is solving polynomials equation in a. If there are any a values for which they are both torsion, we compute the rank of $E_{24}(a)$.

Theorem 6.3. *Let S be the set of a values from Theorem 1.3 for which $\bar{\kappa}(a) = 4$. Let K be a quadratic number field. Then, $S \cap K = \varnothing$.*

Proof. Direction computation. □

Full trees of preimages. We can find an a value with arbitrarily many $\mathbb{Q}$-rational preimages by taking a to be the n-th forward image of any wandering $\mathbb{Q}$-rational point. This gives a very deep but potentially sparse preimage tree. Consequently, one may ask if you can find an a and c which gives a full tree to some level. Clearly, if you allow $K/\mathbb{Q}$ to be of large degree, the answer is any level, so we address this question over $\mathbb{Q}$. For example, here is a list of (c, a) with a 246 preimage arrangement.

$$\left(-\tfrac{5248}{2025}, \tfrac{726745984}{284765625}\right), \quad \left(-\tfrac{17536}{5625}, \tfrac{878382976}{244140625}\right), \quad \left(-\tfrac{9153}{6400}, -\tfrac{437896611}{400000000}\right), \quad \left(-\tfrac{24361}{14400}, -\tfrac{42}{25}\right),$$
$$\left(-\tfrac{20817}{25600}, -\tfrac{1078371711}{6400000000}\right), \quad \left(-\tfrac{180625}{97344}, \tfrac{2845625}{5483712}\right), \quad \left(-\tfrac{158848}{99225}, \tfrac{20844352384}{683722265625}\right).$$

Remark 6.4. We were unable to find any pairs (c, a) over $\mathbb{Q}$ with the full 248 arrangement, but it seems reasonable to expect that such an arrangement exists. We searched by choosing the smallest third preimage having height at most $\log 30{,}000$, since choosing two third preimages which map to same second preimage (up to sign) fixes a unique c value and, hence, a unique a value.

References

[Arslan and Sertöz 1998] F. Arslan and S. Sertöz, "Genus calculations of complete intersections", *Comm. Algebra* **26**:8 (1998), 2463–2471. MR 99e:14058

[Brieskorn and Knörrer 1986] E. Brieskorn and H. Knörrer, *Plane algebraic curves*, Birkhäuser, Basel, 1986. MR 88a:14001 Zbl 0588.14019

[Faber et al. 2009] X. Faber, B. Hutz, P. Ingram, R. Jones, M. Manes, T. J. Tucker, and M. E. Zieve, "Uniform bounds on pre-images under quadratic dynamical systems", *Math. Res. Lett.* **16**:1 (2009), 87–101. MR 2009m:11095 Zbl 1222.11086

[Faber et al. 2011] X. Faber, B. Hutz, and M. Stoll, "On the number of rational iterated pre-images of the origin under quadratic dynamical systems", *Internat. J. Number Theory* **7** (2011), 1781–1806.

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

[Hirzebruch 1966] F. Hirzebruch, *Topological methods in algebraic geometry*, 3rd ed., Grundlehren der Math. Wiss. **131**, Springer, New York, 1966. MR 34 #2573 Zbl 0138.42001

[Hyde 2010] T. Hyde, "On the number of rational iterated pre-images of -1 under quadratic dynamical systems", *Amer. J. Undergradute Res.* **9**:1 (2010), 19–26.

[Kamienny 1992] S. Kamienny, "Torsion points on elliptic curves and q-coefficients of modular forms", *Invent. Math.* **109**:2 (1992), 221–229. MR 93h:11054 Zbl 0773.14016

[Kenku and Momose 1988] M. A. Kenku and F. Momose, "Torsion points on elliptic curves defined over quadratic fields", *Nagoya Math. J.* **109** (1988), 125–149. MR 89c:11091 Zbl 0647.14020

[Masser and Zannier 2008] D. Masser and U. Zannier, "Torsion anomalous points and families of elliptic curves", *C. R. Math. Acad. Sci. Paris* **346**:9-10 (2008), 491–494. MR 2009j:11089 Zbl 1197.11066

[Masser and Zannier 2012] D. Masser and U. Zannier, "Torsion points on families of squares of elliptic curves.", *Math. Ann.* **352**:2 (2012), 453–484. Zbl 06006368

[Mazur 1977] B. Mazur, "Modular curves and the Eisenstein ideal", *Inst. Hautes Études Sci. Publ. Math.* 47 (1977), 33–186. MR 80c:14015 Zbl 0394.14008

[Oguiso and Shioda 1991] K. Oguiso and T. Shioda, "The Mordell–Weil lattice of a rational elliptic surface", *Comment. Math. Univ. St. Paul.* **40**:1 (1991), 83–99. MR 92g:14036 Zbl 0757.14011

[Sendra et al. 2008] J. R. Sendra, F. Winkler, and S. Pérez-Díaz, *Rational algebraic curves: A computer algebra approach*, Algorithms and Computation in Mathematics **22**, Springer, Berlin, 2008. MR 2009a:14073 Zbl 1129.14083

[Shioda 1990] T. Shioda, "On the Mordell–Weil lattices", *Comment. Math. Univ. St. Paul.* **39**:2 (1990), 211–240. MR 91m:14056 Zbl 0725.14017

Received: 2010-09-13 Revised: 2011-05-14 Accepted: 2011-05-15

bhutz@gc.cuny.edu
Department of Mathematics, CUNY Graduate Center,
365 Fifth Ave, New York, NY 10016, United States

thyde12@amherst.edu
Department of Mathematics and Computer Science,
Amherst College, Amherst, MA 01002, United States

benkrause23@math.ucla.edu
Department of Mathematics, University of California,
Box 951555, Los Angeles, CA 90095-1555, United States

The Steiner problem on the regular tetrahedron

Kyra Moon, Gina Shero and Denise Halverson

(Communicated by Frank Morgan)

The Steiner problem involves finding a shortest path network connecting a specified set of points. In this paper, we examine the Steiner problem for three points on the surface of a regular tetrahedron. We prove several important properties about Steiner minimal trees on a regular tetrahedron. There are infinitely many ways to connect three points on a tetrahedron, so we present a way to eliminate all but a finite number of possible solutions. We provide an algorithm for finding a shortest network connecting three given points on a regular tetrahedron. The solution can be found by direct measurement of the remaining possible Steiner trees.

1. Introduction

The *Steiner problem* asks to find a shortest path network to connect a given set of points on a surface. In this paper we will study the three point Steiner problem on a regular tetrahedron. We will provide an algorithm in Section 10, Algorithm 10.1, that determines a solution to the three point Steiner problem on the regular tetrahedron.

On the Euclidean plane, the Steiner problem has been studied extensively; see [Gilbert and Pollak 1968; Hwang et al. 1992; Ivanov and Tuzhilin 1994, Chapter 9; Melzak 1961; Zacharias 1914–1921]. The Steiner problem for three points on the Euclidean plane was formally introduced in the seventeenth century by Fermat; see [Hwang et al. 1992; Kuhn 1974; Zacharias 1914–1921]. A general algorithm to find the solution to the Steiner problem for n points on the Euclidean plane was first developed by Melzak [1961] (see also [Hwang et al. 1992]).

The Steiner problem on the surface of the tetrahedron is not as straightforward as on the plane. In particular, a geodesic segment connecting any two points on the surface of the tetrahedron is not unique (see top part of Figure 1 on next page). Consequently, there are infinitely many locally stable shortest-length trees connecting any three points on the surface of the tetrahedron (see Figure 1, bottom). In this

MSC2010: primary 05C05; secondary 51M15.

Keywords: Steiner problem, length minimization, regular tetrahedron, piecewise-linear surface.

Funded by NSF grant DMS-0755422 and a BYU undergraduate research grant.

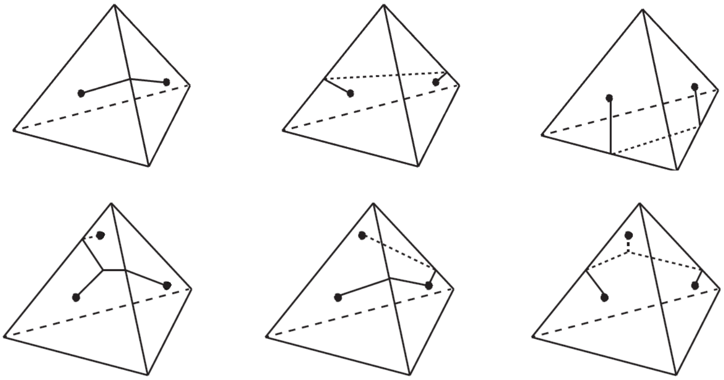


Figure 1. Candidates for a shortest path (top) and for a shortest tree (bottom).

paper, we provide an algorithm that eliminates all but a small number of path networks that need be considered as possible minimizers. Amongst these remaining candidates, a shortest path network can be found using direct measurement.

This research contributes to the growing set of strategies for solving Steiner problems on surfaces in general. Algorithms exist to find the solution for the Steiner problem on certain surfaces of constant curvature. The problem was studied in [Weng 2001; Litwhiler and Aly 1980; Brazil et al. 1998] for on curved surfaces, including spheres. March and Halverson [2005] studied Steiner trees in hyperbolic space. Lee et al. [2011] studied the Steiner problem on wide and narrow cones. Penrod [2007] and May and Mitchell [2007] developed algorithms to solve Steiner problems on the flat torus. Caffarelli et al. [2010] studied the Steiner problem on surfaces of revolution. Brune and Sipe [2009] developed an algorithm to find a shortest path between two points on the surface of the regular tetrahedron. This research about the Steiner problem on the regular tetrahedron may provide further insight into the Steiner problem on more general piecewise linear surfaces.

2. Preliminaries

We begin by setting up the basic framework for the Steiner problem on a regular tetrahedron $\mathcal{T}$. Let $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$ be a set of given points on $\mathcal{T}$ called *terminal points*, and let L be a path network (also on $\mathcal{T}$) connecting the points in $\mathcal{A}$. A *path network* connects a collection of arcs, only possibly meeting at the endpoints such that the network contains a path connecting any two points of $\mathcal{A}$. If L is a shortest path network, the edges must be geodesics. L must also be a tree since if L contained a cycle, one of the edges could be removed. The goal of the Steiner problem is to find a shortest path network L connecting the points of $\mathcal{A}$. A shortest path network may have additional vertices called *Steiner points*. The solution to the Steiner problem is called the *Steiner minimal tree*, which will be denoted by SMT($\mathscr{A}$).

As defined in [Hwang and Weng 1986], a tree with n fixed points is called a *Steiner tree* on n fixed points if it satisfies the following conditions:

(1) There are at most $n-2$ Steiner points.

(2) Each Steiner point has exactly three incident edges.

(3) Any pair of edges meeting at any vertex of the tree form an angle with measure at least 120°.

Note that for a tree with no degree-two Steiner points, the number of edges minus the number of vertices is 1, which in fact implies condition 1. A tree that has exactly $n-2$ Steiner points is called a *full Steiner tree*. A tree that has fewer than $n-2$ Steiner points is called a *degenerate Steiner tree*.

The Steiner problem for n fixed points on the plane can be solved in finite time using Melzak's algorithm [1961]. We will utilize these results for the regular tetrahedron since the plane can be viewed as a branched cover of the regular tetrahedron. The Steiner problem on $\mathcal{T}$ is more complex than on the plane because there are infinitely many geodesics that could connect two points. Thus, the process of solving the Steiner problem on $\mathcal{T}$ is initially a problem of narrowing down potential path networks.

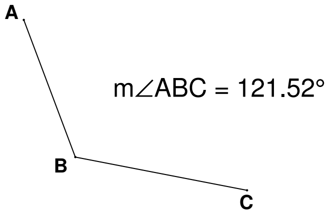
The algorithm used to solve the 3-point Steiner problem in Euclidean space was developed by Torricelli, Cavalieri, Simpson, Heinen, and Bertrand (see [Hwang et al. 1992]). For convenience, we repeat it here.

Algorithm 2.1. This algorithm provides a shortest network connecting three given points in Euclidean space.

(1) Let A, B, and C be given. Label A, B, and C so that $m\angle ABC \geq m\angle ACB$ and $m\angle ABC \geq m\angle BAC$.

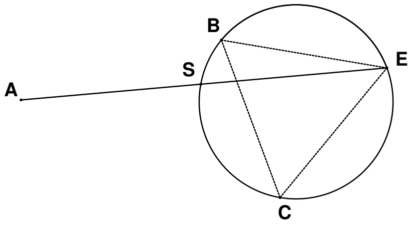
(2) Determine whether Case 1 or 2 applies.

Case 1. If $m\angle ABC > 120°$, the Steiner minimal tree is degenerate and it is $\overline{AB} \cup \overline{BC}$. The algorithm is complete (see figure for example).



Case 2. If $m\angle ABC \leq 120°$, proceed to Steps (3)–(6).

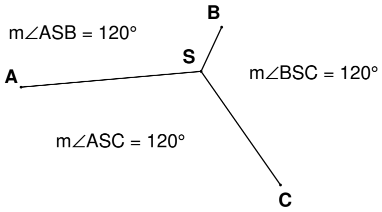
(3) Create an equilateral triangle $\triangle BCE$ where E is on the opposite side of $\overleftrightarrow{BC}$ from A.



(4) Construct $\overline{EA}$. This line segment is called the Simpson line. (The length of the Simpson Line is known to have the same length as the SMT(A, B, C) [Hwang et al. 1992].)

(5) Next, circumscribe a circle about $\triangle BCE$. The point of intersection of that circle and $\overline{EA}$ is the Steiner point S.

(6) Connect each of A, B, and C to S to form SMT(A, B, C). By construction, every two edges of the tree which meet at the Steiner point have angle 120° [Gilbert and Pollak 1968]. The algorithm is complete.

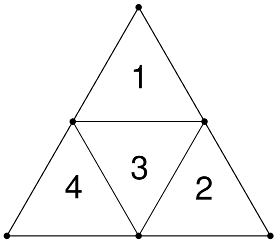


Another observation relevant to our discussion of the Steiner problem on the regular tetrahedron is that no geodesic passes through the vertices of a narrow cone [Lee et al. 2011]. Since a small neighborhood of a vertex is a narrow cone, no shortest path network will pass through any vertices of $\mathcal{T}$. Hence, a shortest path network can only meet a vertex of $\mathcal{T}$ if a fixed point is placed on that vertex [Ivanov and Tuzhilin 1994, Chapter 9].

3. Tiling the plane

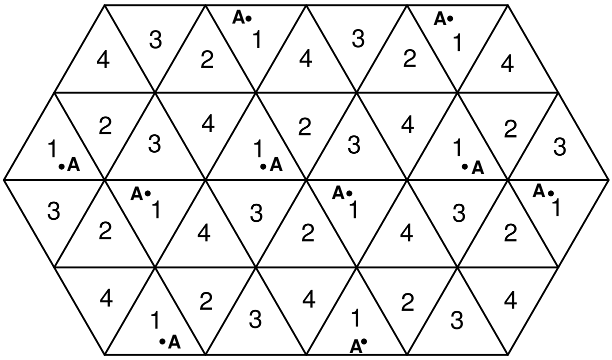
In this section we will show how to construct a branched covering of the plane onto the regular tetrahedron. For further reference, see [Ivanov and Tuzhilin 1994, Chapter 9].

Consider a regular tetrahedron with faces labeled 1, 2, 3, and 4. Cut along the edges common to faces 1 and 2, 1 and 4, and 2 and 4 and lay it on the plane, as shown in the figure. We will use this configuration to tile the plane.

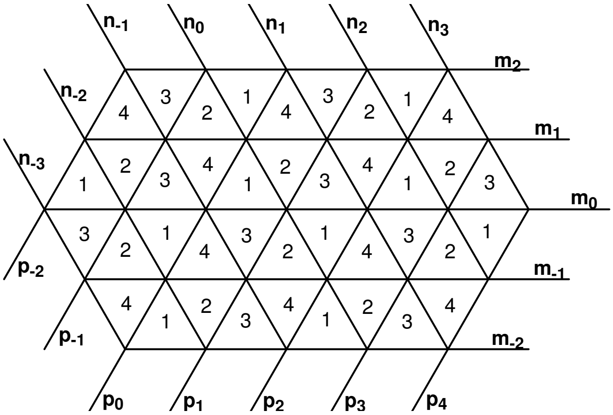


Notice that face 1 is adjacent to face 2 on the tetrahedron. Thus, in order to represent that on the plane, we must place a tile corresponding to face 2 so it becomes adjacent to a tile corresponding to face 1. This is accomplished by placing a tile corresponding to face 2 that is an 180° rotation about a common vertex. Similarly, we must place a tile corresponding to face 4 so that the tile corresponding to face 1 and a tile corresponding to face 4 have a common edge in the plane as they do on the tetrahedron. Since each face on the tetrahedron is adjacent to the other three faces, then each face should be adjacent to all of the other faces on the plane. If copies of each face are placed at 180° rotations about each of their respective vertices, this results in a comprehensive tiling of the Euclidean plane.

Points on $\mathcal{T}$ will be represented by lower case letters. The corresponding points in the tiling will be represented by corresponding capital letters. Assume a is on face 1 on $\mathcal{T}$. Then for each tile corresponding to face 1, there is a copy of A on the tile. Two adjacent tiles contain copies of A which are 180° rotations about the common vertex of the tile containing A. A small section of the tiling can be seen here:


4 3 2 A• 1 4 3 2 A• 1 4
1 •A 2 3 4 1 •A 2 3 4 1 •A 2 3
3 2 A• 1 4 3 2 A• 1 4 3 2 A• 1
4 1 •A 2 3 4 1 A• 2 3 4

We introduce a coordinate system to notate the different faces of the tiling. In the tiling, the horizontal lines that separate the triangles will be known as m_i, for $i = \ldots, -2, -1, 0, 1, 2, \ldots$. Similarly define n_i as the lines with the slope equal to $-\sqrt{3}$. Finally define p_i as the lines with slope $\sqrt{3}$. We thus obtain the following arrangement:



Using this coordinate system, we can identify individual tiles. For any tile that is bounded by m_x, n_y, and p_z, we will denote it as $T_{(x,y,z)}$. Without loss of generality, we will assume that $T_{(0,0,0)}$ corresponds to face 1, $T_{(1,-1,0)}$ corresponds to face 2, $T_{(0,-1,1)}$ corresponds to face 3, and $T_{(1,0,1)}$ corresponds to face 4. Though each face of the tetrahedron is replicated infinitely many times, each tile in the tiling has a unique labeling according to the lines that bound it.

We now show that this tiling is a branched covering of the plane onto the regular tetrahedron. Let $\Pi : \mathbb{R}^2 \to \mathcal{T}$ be the natural continuous map that takes each tile of the plane to its corresponding face in $\mathcal{T}$ homeomorphically. Let $\mathcal{V}$ be the vertex set of $\mathcal{T}$. Note that Π is a branched covering map with branch set $\mathcal{V}$. Then the map

$$\pi : \mathbb{R}^2 - \Pi^{-1}(V) \to \mathcal{T} - \mathcal{V}$$

(which is a restriction of Π) is a covering map of $\mathcal{T} - \mathcal{V}$. Since π is a covering map, it has the following lifting property: Suppose $a \in \mathcal{T} - \mathcal{V}$ and $A \in \Pi^{-1}(a)$. Then any path $\alpha : [0, 1] \to \mathcal{T} - \mathcal{V}$ so that $\alpha(0) = a$ has a unique lift to a path $\tilde{\alpha} : [0, 1] \to \mathbb{R}^2 - \Pi^{-1}(\mathcal{V})$ with $\tilde{\alpha}(a) = A$. The map $\tilde{\alpha}$ is a lift in the sense that $\pi \circ \tilde{\alpha} = \alpha$. It follows that any embedded path network in $\mathcal{T} - \mathcal{V}$ containing a can be uniquely lifted to a path network containing A.

Note that in the case that $a \in \mathcal{V}$ and $\Pi(A) = a$, for any embedded path network containing a in $\mathcal{T}$, there are two lifts of the path network containing A. These lifts are 180° rotations of each other about A.

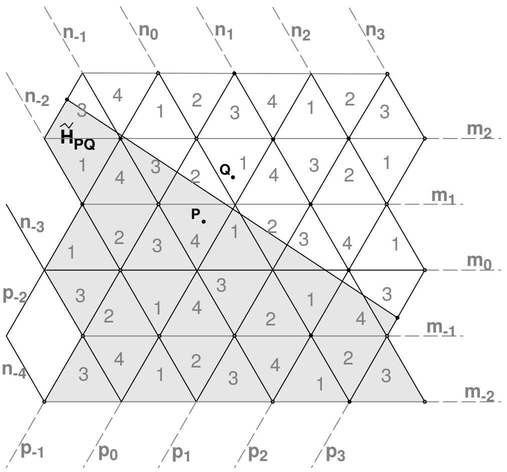
4. The two point problem

This section will briefly describe an algorithm used to construct a shortest path between any two points on a regular tetrahedron. For further details on this process, refer to [Brune and Sipe 2009]. The algorithm detailed here will depend heavily on the following basic geometric property:

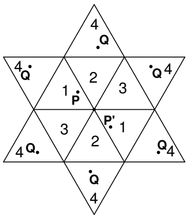
Property 4.1. *Given any two points A and B on the plane, construct the perpendicular bisector of $\overline{AB}$ and call it P_{AB}. If X is on the A side of P_{AB}, then X is closer to A. If X is on the B side of P_{AB}, then X is closer to B.*

Definition 4.2. Given two points P and Q on the plane, define $\tilde{H}_{PQ}$ to be the half-plane cut by the perpendicular bisector of P and Q on the P side; that is,

$$\tilde{H}_{PQ} = \{X \mid PX \leq QX\}.$$



The algorithm: a brief synopsis. Suppose there are two points p and q on distinct faces of the tetrahedron. Suppose $\mathbb{R}^2$ is tiled as in Section 3. Recall that $\Pi : \mathbb{R}^2 \to \mathcal{T}$ is the covering map and $\mathbb{R}^2$ is tiled as in Section 3. Then $\Pi^{-1}(p)$ and $\Pi^{-1}(q)$ contain infinitely many points. Let $P \in \Pi^{-1}(p)$. We want to find a point $Q \in \Pi^{-1}(q)$ that realizes a shortest path from p to q. The points of $\Pi^{-1}(q)$ that could realize a shortest path to P can be restricted to a star-shaped region. The region consists of an interior hexagon which contains the point P, outlined by six tiles which contains points of $\Pi^{-1}(q)$. This region is called an *i-star* for $i = 1, 2, 3,$ or 4, where i is the face of the tetrahedron containing q. We illustrate a 4-star when p is on face 1 and q is on face 4:



It was proved in [Brune and Sipe 2009] that this i-star always contains a shortest path between two points.

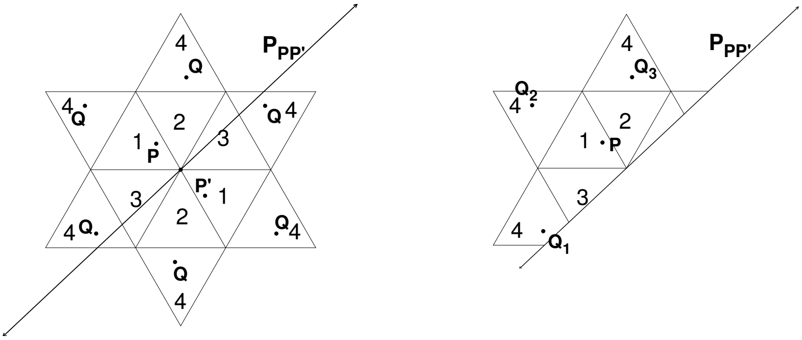
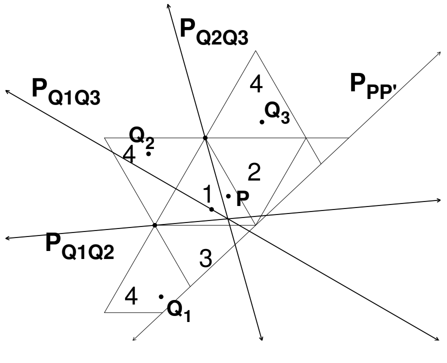


Figure 2. Reducing the number of possible points of $\Pi^{-1}(q)$ that can realize a shortest path.

There is a cutting technique that has been shown to reduce the number of possible points of $\Pi^{-1}(q)$ that could realize a shortest path. Begin by constructing the line segment from point P to the point $P' \in \Pi^{-1}(p)$, also located within the 4-star. Then, construct the perpendicular bisector of $\overline{PP'}$, denoted $P_{PP'}$ (see Figure 2, left). Every point of $\Pi^{-1}(q)$ within the star that falls on the same side of l as P will now be the only copies of $\Pi^{-1}(q)$ considered for the shortest path. The portion of the star-shaped region which is on the P side of $P_{PP'}$ is called τ (see Figure 2, right).

There are three points of $\Pi^{-1}(q)$ in τ which we will label as Q_1, Q_2, and Q_3, as shown in Figure 2, right. (If $P_{PP'}$ contains a point of $\Pi^{-1}(q)$ in τ, then it contains another point of $\Pi^{-1}(q)$ and either point in $\Pi^{-1}(q)$ in τ can be discarded.) To find $\min\{PQ_i\}$ where $i = 1, 2, 3$, we construct $\tilde{H}_{Q_iQ_j}$ for $i = 1, 2, 3$ and $j \neq i$.

Note that the boundary of $\tilde{H}_{Q_iQ_j}$ is $P_{Q_iQ_j}$. If Q_i is closest to P, then P must lie in $\tilde{H}_{Q_iQ_j} \cap \tilde{H}_{Q_iQ_k}$. Note that if P is equally close to Q_i and Q_j, then P lies in both $\tilde{H}_{Q_iQ_j} \cap \tilde{H}_{Q_iQ_k}$ and $\tilde{H}_{Q_jQ_i} \cap \tilde{H}_{Q_jQ_k}$. In the figure below, a shortest path is realized by $\overline{PQ_3}$. Hence, P lies in $\tilde{H}_{Q_3Q_1} \cap \tilde{H}_{Q_3Q_2}$. In particular, $\Pi(\overline{PQ_3})$ is the minimal geodesic connecting p and q and will traverse faces 1, 2, and 4.



5. Overview

Suppose $\{x, y, z\} \in \mathcal{T}$. Recall that Π is the branched covering map described in Section 3. Thus $\Pi^{-1}(x)$, $\Pi^{-1}(y)$, and $\Pi^{-1}(z)$ contain infinitely many points. Hence, there are also infinitely many distinct Steiner trees connecting points x, y and z. Our goal in this paper is to narrow down the number of combinations in the tiled plane which may realize the solution.

As stated earlier, we will divide our discussion of this problem into three cases:

Case 1: Three points that can be considered to be on one face of $\mathcal{T}$.

Case 2: Three points that can be considered to be on three distinct faces of $\mathcal{T}$.

Case 3: Any configuration of three points that does not fit into the first two cases (i.e., three points that can only be considered to be on two distinct faces).

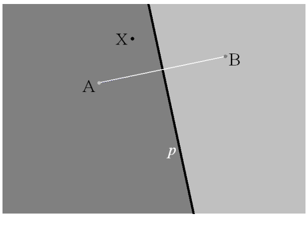
Section 6 will address the simplest case where all three points are on a common face of the tetrahedron. Section 7 will introduce the strategies needed for Sections 8 and 9. In Section 8, we will discuss case 2, and in Section 9 we will discuss case 3. We will discuss how to solve the problem for any specific positioning of the points in Section 10.

6. Case 1: Three points on one face

We know by a theorem proved in [Brune and Sipe 2009] that a shortest path network connecting n points contained on the same face of a regular tetrahedron is contained within that face. Thus, the Steiner minimal tree for three points on the same face of a tetrahedron can be constructed in that face using the algorithm described in Algorithm 2.1.

7. Geometric properties of Steiner minimal trees

Given $a, b, c \in \mathcal{T}$ and the corresponding point sets on the tiled plane, there are many ways that points can be selected, each corresponding to a Steiner tree on $\mathcal{T}$. However, only certain of the combinations realize the Steiner minimal tree on the tetrahedron. The next several results represent strategies that help eliminate fruitless combinations. At this point the reader is encouraged to reread Property 4.1, describing the situation illustrated here:



Lemma 7.1 (perpendicular bisector rule I). *Suppose A, $A' \in \Pi^{-1}(a)$ such that A is on tile T and A' is on tile T'. Then for any point B on T, $AB \le A'B$. If neither A nor B are a common vertex of T and T', then $AB < A'B$.*

Proof. Let $b = \Pi(B)$. Note that a and b are on the same face. We know from a theorem proved in [Brune and Sipe 2009] that a shortest path network connecting n points on the same face is in that same face and here is $\overline{ab}$, which is realized by $\overline{AB}$ in T. Since $\overline{AB}$ is a minimum of all paths $A'B$ where $A' \in \Pi^{-1}(a)$, then for all $A' \neq A$, $AB \le A'B$. If A is not a common vertex of T and T', then $A \neq A'$, so $P_{AA'}$ is defined. If B is not a common vertex of T, then $B \in P_{AA'}$. Thus $AB < A'B$. □

Next, let A, B, and C be points in the tiled plane such that

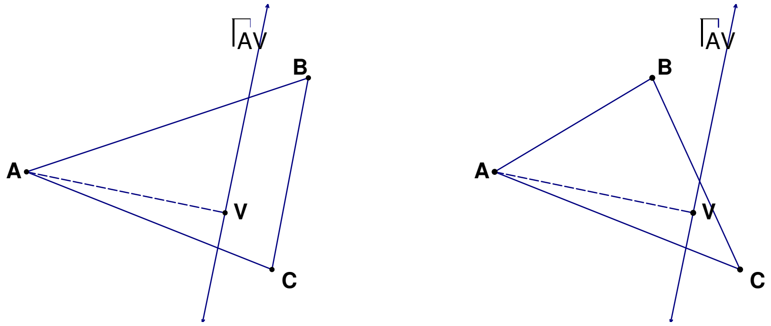
$$\Pi(\mathrm{SMT}(A, B, C)) = \mathrm{SMT}(a, b, c).$$

We will show that the convex hull of the triangular region formed from A, B, and C cannot contain a vertex of the tiled plane unless that vertex is one of A, B, or C. However, before we prove this, we introduce a definition and a property of triangular regions in general.

Definition 7.2. Given two points X and V, let Γ_{XV} be the line perpendicular to $\overline{XV}$ through V.

Lemma 7.3. *Suppose there is a triangular region with vertices A, B, and C that contains the point V in the interior. Then there is an $X \in \{A, B, C\}$ such that Γ_{XV} separates X from $\{A, B, C\} - \{X\}$.*

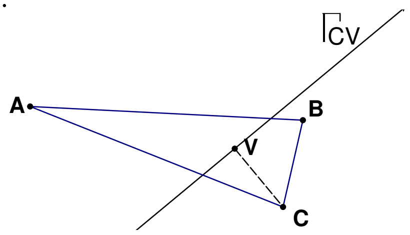
Proof. If Γ_{AV} separates A from BC, the proof is done (left figure):



Otherwise, one of B or C is on the same side of Γ_{AV} as A.

Without loss of generality, suppose B is on the same side of Γ_{AV} as A (right figure). Then $m\angle AVB \le 90°$. Then if Γ_{CV} separates C from A and B, the proof is done.

If not, one of A or B is on the same side of Γ_{CV} as C. In the former case we have $m\angle CVA \le 90°$, while in the latter we have $m\angle CVB \le 90°$. Here is an illustration of the second possibility:



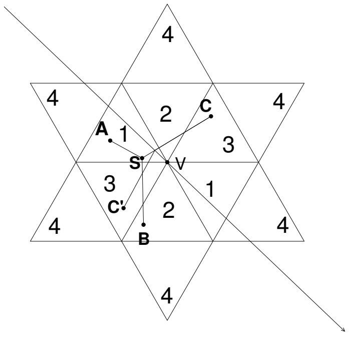
Thus, either $m\angle AVC + m\angle AVB \leq 180°$ or $m\angle CVB + m\angle AVB \leq 180°$. In either case, we are in contradiction with the hypothesis that V is in the interior of $\triangle ABC$. Thus, there exists an $X \in \{A, B, C\}$ such that Γ_{XV} separates X from $\{A, B, C\} - \{X\}$. □

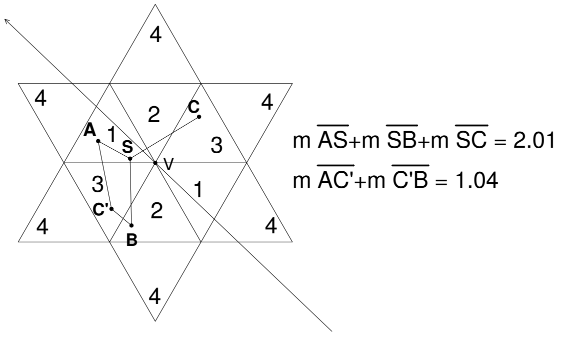
Theorem 7.4 (vertex rule). *Suppose a, b, and $c \in \mathcal{T}$ and*

$$\Pi(\mathrm{SMT}(A, B, C)) = \mathrm{SMT}(a, b, c).$$

Then the image of the convex hull of $\triangle ABC$ under Π cannot contain a vertex v, unless v is one of a, b, or c.

Proof. By way of contradiction, suppose a vertex V of the tiling is contained in the interior of the convex hull of $\triangle ABC$. Construct $\mathrm{SMT}(A, B, C)$, and label the Steiner point S (the Steiner tree may possibly be degenerate). Using Lemma 7.3, we may assume without loss of generality that Γ_{CV} separates C from both A and B. Reflect the part of the path on the C side of Γ_{CV} across Γ_{CV}. Let C' be the reflection of C across Γ_{CV}. Note that the partially reflected path connects A, B, and C' and is equal in length to $\mathrm{SMT}(A, B, C)$. Thus, there is an alternate choice of points in $\Pi^{-1}(a)$, $\Pi^{-1}(b)$, and $\Pi^{-1}(c)$ which is at least as short as $\mathrm{SMT}(A, B, C)$. If S is on the opposite side of Γ_{CV} as C, we can shorten the tree by replacing SC with SC' (see figure on the left). If S is on the same side of Γ_{CV} as C, we can shorten the tree by replacing SA with SA' and SB with SB', where A' and B' are the reflections of A and B across Γ_{CV}, respectively. If S is on Γ_{CV}, then $SC = SC'$, so either tree is the same length. However, the tree containing A, B,




4
4
2
C
4
A
1
S
3
V
3
1
C'
2
4
B
4
4
m $\overline{AS}$+m $\overline{SB}$+m $\overline{SC}$ = 2.01
m $\overline{AC'}$+m $\overline{C'B}$ = 1.04

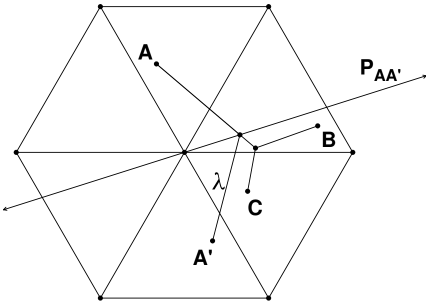
and C' will no longer meet the 120° condition for Steiner trees, and will not be $\mathrm{SMT}(A, B, C')$. Thus $\mathscr{L}(\mathrm{SMT}(A, B, C')) < \mathscr{L}(\mathrm{SMT}(A, B, C))$, which implies that $\Pi(\mathrm{SMT}(A, B, C)) \neq \mathrm{SMT}(a, b, c)$. □

Theorem 7.5 (perpendicular bisector rule II). *Let $A, A' \in \Pi^{-1}(a)$ on the tiled plane be distinct. If $P_{AA'}$ separates $\{B, C\}$ from A, then*

$$\mathscr{L}(\mathrm{SMT}(A', B, C)) < \mathscr{L}(\mathrm{SMT}(A, B, C)).$$

Hence, $\Pi(\mathrm{SMT}(A, B, C)) \neq \mathrm{SMT}(a, b, c)$.

Proof. Let λ be the reflection of the part of $\mathrm{SMT}(A, B, C)$ on the A side of $P_{AA'}$ across $P_{AA'}$:

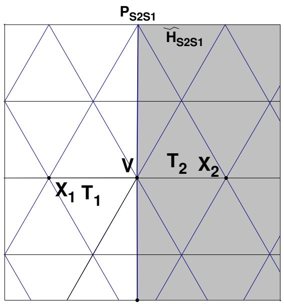


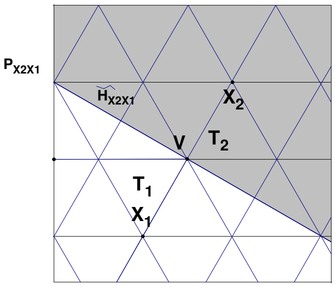
Note that λ uses the point A' as a terminal, thus it is a path network connecting A', B, and C. By a similar argument as in Theorem 7.4, we obtain

$$\mathscr{L}(\mathrm{SMT}(A, B, C)) = \mathscr{L}(\lambda) > \mathscr{L}(\mathrm{SMT}(A, B, C')). \qquad \square$$

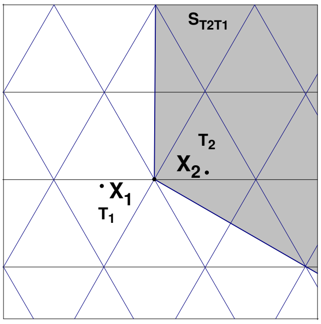
Sectors and half-planes.

Definition 7.6. Fix a vertex V of the tiled plane, and let T_1 and T_2 be tiles (not necessarily adjacent to V) that are mapped to one another with respect to 180° rotation about V. Define the sector $S_{T_2T_1}$ as the intersection of all half-planes $\tilde{H}_{X_2X_1}$, where X_1 runs over all points in T_1 and X_2 is it image under a 180° rotation about V. Clearly $\tilde{H}_{X_2X_1}$ is fully determined by the direction of the vector VX_1; thus by considering two extreme cases for this direction, as here:



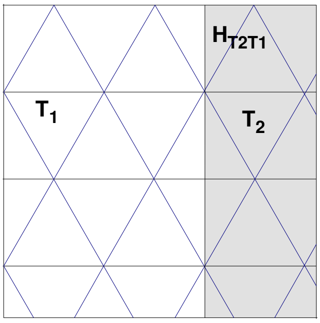


we conclude that the $S_{T_2T_1}$ is the intersection of the half-planes $\tilde{H}_{X_2X_1}$ obtained in these two cases:



Next, if Y and Z are arbitrary points belonging to tiles T_1 and T_2, respectively, we set $S_{YZ} = S_{T_2T_1}$.

Definition 7.7. Let T_1, T_2 be tiles that are translates of each other on the tiled plane, satisfying $\Pi(T_1) = \Pi(T_2)$. Then the intersection of all half-planes $\tilde{H}_{X_2X_1}$ where $X_i \in T_i$ and $\Pi(X_1) = \Pi(X_2)$, is denoted by $H_{T_2T_1}$.



If Y and Z are arbitrary points belonging to tiles T_1 and T_2, respectively, we set $S_{YZ} = S_{T_2T_1}$.

Theorem 7.8 (Steiner point rule). *Let A, B, and C be points in the tiled plane such that $\Pi(\mathrm{SMT}(A, B, C))$ is a Steiner minimal tree on the tetrahedron. Suppose that S is the Steiner point of* SMT(A, B, C). *If S' is any other point of* $\Pi^{-1}(\Pi(s))$, *then $XS \leq XS'$ for $X = A$, B, and C.*

Proof. Without loss of generality, assume that $X = C$. By way of contradiction, suppose $CS' < CS$. Then there exists a point $C' \in \Pi^{-1}(c)$ such that $CS' = C'S$. This implies that

$$\begin{aligned}\mathcal{L}(\mathrm{SMT}(A, B, C)) &= AS + BS + CS > AS + BS + C'S \\ &\geq \mathcal{L}(\mathrm{SMT}(A, B, C')),\end{aligned}$$

as needed. (See Figure 3 on next page.) □

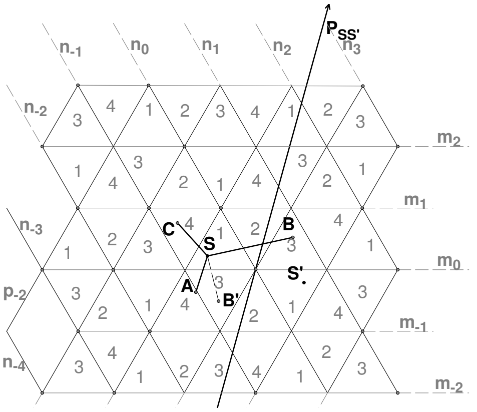


Figure 3. Toward the proof of Theorem 7.8.

8. Case 2: Three points on three distinct faces

When the three points can be viewed to lie on three distinct faces, we use the following procedure to determine the possible configurations of the points on the tiled plane which may realize the Steiner minimal tree. Our arguments apply also when the three points can be viewed to lie on two or one face, as may be the case if one or more of the points lie on vertices or edges. For example, if one point is in the interior of a face, another point is in the interior of another face, and the third point is on a vertex shared by both faces, then we can assign the third point to the third face which shares that vertex, and the configuration is in the realm of Case 2.

Triple ribbon region. Recall the labeling system introduced in Section 3, in which m_i, n_i, and p_i represent the horizontal, negative slope, and positive slope lines, respectively. Also recall that the triangle that is bounded by m_x, n_y, and p_z will be denoted as $T_{(x,y,z)}$.

Let a, b, and c be points on the tetrahedron such that s is the Steiner point for SMT(a, b, c). Let τ_0 be the shaded region in Figure 4. Since τ_0 contains copies of the tiles corresponding to all four faces, a copy of $S \in \Pi^{-1}(s)$ must lie within τ_0.

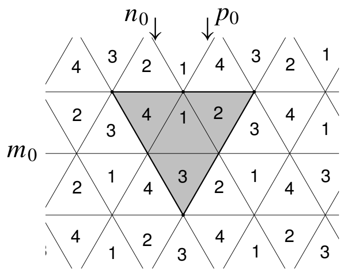
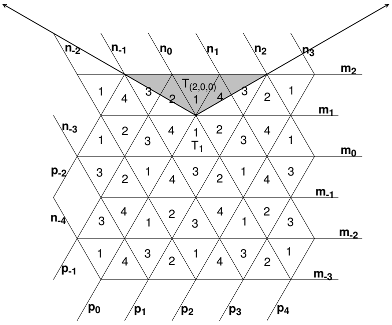


Figure 4. The region τ_0.

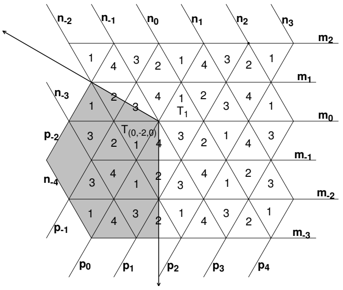
Let $\mathcal{S}^* = \Pi^{-1}(\Pi(s)) - \{S\}$. We will determine a region $\mathcal{R}$ such that given a point $P \in \mathcal{R}$, $PS \leq PS'$ for any $S' \in \mathcal{S}^*$. It follows from Theorem 7.8 that any points not in $\mathcal{R}$ cannot be the fixed points of the Steiner minimal tree that contains S and realizes $\mathrm{SMT}(a, b, c)$.

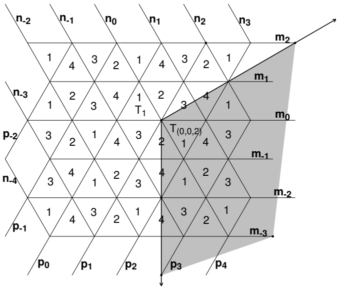
In order to simplify the process, we will first determine the region $\mathcal{R}_i$ that contains all points closer to T_i than to any other tile corresponding to face 1. Then $\mathcal{R} = \bigcup \mathcal{R}_i$. We will call $\mathcal{R} = \bigcup \mathcal{R}_i$ the *triple ribbon region.*

Reductions. Let $i = 1$. Let S' be the 180° rotation of S about the vertex $V = T_1 \cap T_{(2,0,0)}$. Then any point $X \in S_{T_{(2,0,0)}T_1}$ is closer to S' than S. Thus no fixed point is in $S_{T_{(2,0,0)}T_1}$:

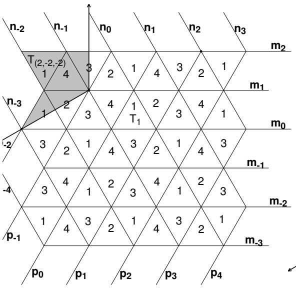


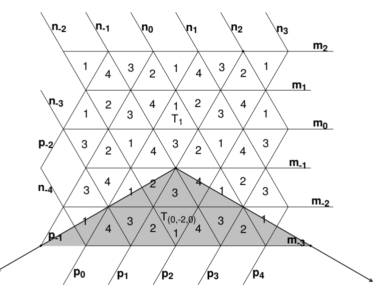
Likewise, no fixed points will be found in $S_{T_{(0,-2,0)}T_1}$ or $S_{T_{(0,0,2)}T_1}$:

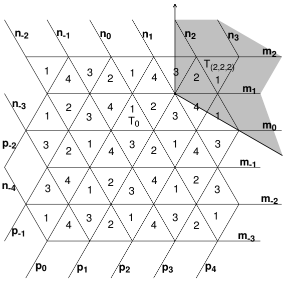




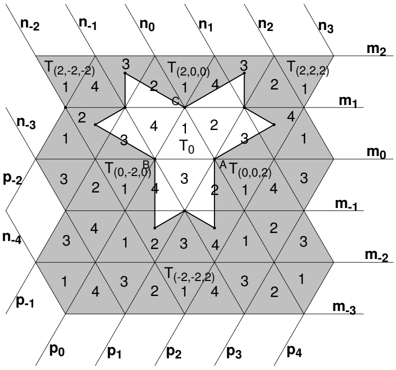
There are also no fixed points to be found in $S_{T_{(2,-2,-2)}T_1}$, $S_{T_{(-2,-2,2)}T_1}$, and $S_{T_{(2,2,2)}T_1}$:







R_1 is the closure of the region remaining when the shaded regions in the six figures of the previous page are cut away. It is shown in white here:



Regions R_2, R_3, and R_4 are found similarly. The union of all these regions, $\mathcal{R} = \bigcup_{i=1}^{4} \mathcal{R}_i$, is the triple ribbon region (Figure 5).

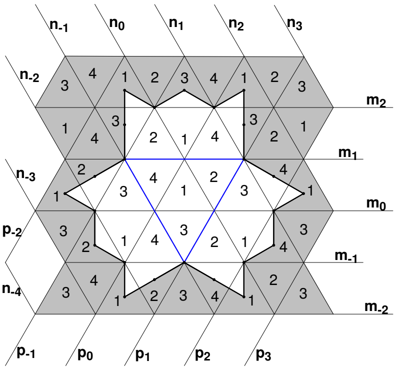


Figure 5. The triple ribbon region (in white).

Regardless of the location of s on the tetrahedron, a copy of $\Pi^{-1}(\mathrm{SMT}(a, b, c))$ is contained within the triple ribbon region. Thus, it is sufficient to check only the combinations of fixed points in the triple ribbon region.

Although the number of potential path networks needed to be checked to find $\mathrm{SMT}(a, b, c)$ is a finite number, it is still a significant number. Note that there are six tiles meeting the triple ribbon region corresponding to face i for $i = 2, 3, 4$. Thus there are $6 \times 6 \times 6 = 216$ combinations to consider given the specification of points in certain faces of τ_0. Hence, we continue to make further reductions.

Horn removal. We subdivide the triple ribbon region as follows. The closure of the bounded white region in Figure 6 (on the next page) is called the *badge region*. The small black triangles, which make up the difference between the triple ribbon region and the badge region, are called the *horns*.

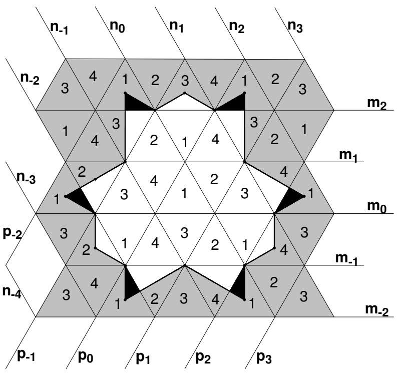


Figure 6. The badge region (closure of the polygon in white) and the horns (in black).

Proposition 8.1. *Suppose a, b, and c are three points on distinct faces of $\mathcal{T}$, none of which are chosen to be face* 1. *Then there is a copy of*

$$\mathrm{SMT}(A, B, C) \in \Pi^{-1}(\mathrm{SMT}(a, b, c))$$

on the tiled plane which is contained in the badge region centered about a tile corresponding to face 1 *with Steiner point S contained in the triangular region τ_0* (see Figure 4).

Proof. Without loss of generality, assume that a is contained on face 3, b is contained on face 4, and c is contained on face 2. Let $A \in \Pi^{-1}(a)$, $B \in \Pi^{-1}(b)$, and $C \in \Pi^{-1}(c)$ lie in the triple ribbon region such that $\Pi(\mathrm{SMT}(A, B, C)) = \mathrm{SMT}(a, b, c)$. Note that no portion of the horns contains any points of $\Pi^{-1}(a)$, $\Pi^{-1}(b)$, or $\Pi^{-1}(c)$ and therefore cannot contain A, B, or C. Let H_1 be the horn bounded by m_2, n_0, and p_{-1} that is outside the badge region.

Suppose an edge of $\mathrm{SMT}(A, B, C)$ meets H_1 outside the badge region. If the interior of an edge passes through either side of the horn not on m_2, the edge must meet the shaded region. But by hypothesis, $\mathrm{SMT}(A, B, C)$ must lie entirely within the triple ribbon region. Thus the edge may only pass through the boundary of the horn on m_2. If so, the only possibility is that one of the endpoints of the edges is contained in H_1. Thus a fixed point is contained in the interior of the horn, and hence contained in the interior of face 1. But by hypothesis, face 1 was not selected as one of the faces containing fixed points. Therefore, an edge of $\mathrm{SMT}(A, B, C)$ does not meet H_1. By a similar argument, $\mathrm{SMT}(A, B, C)$ cannot meet any horn. □

Reduction to the piping region. Using Theorem 7.4 and Theorem 7.5, we will now demonstrate that a lift of the Steiner minimal tree can be contained in a subset of the badge region called the *piping region* (Figure 7). What is left over of the

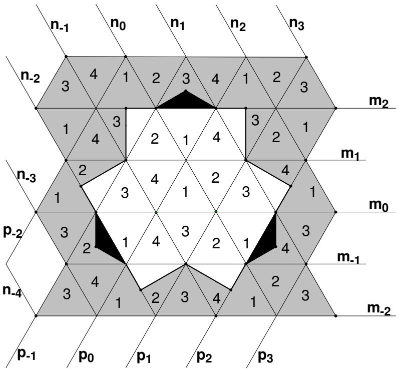


Figure 7. The piping region (closure of the polygon in white) and the flaps (in black).

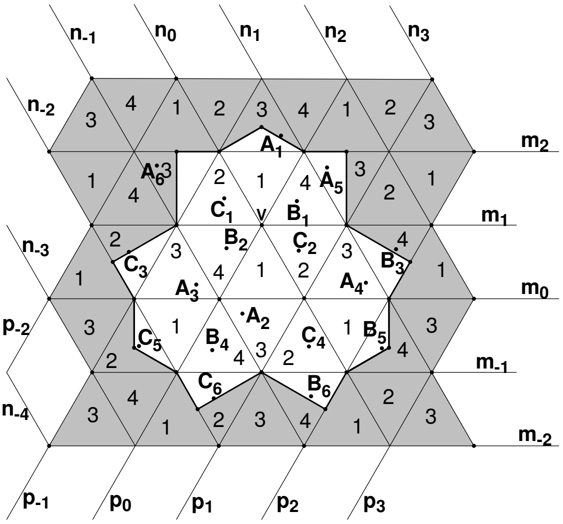
badge region is called the (*top*) *flaps*. We will show that if SMT(A, B, C) realizes SMT(a, b, c) and is contained in the badge region, then SMT(A, B, C) does not meet the flaps outside the piping region.

Theorem 8.2. *Suppose a, b, and c are three points on distinct faces of $\mathcal{T}$, none of which chosen to be face* 1. *Suppose SMT$(A, B, C) \in \Pi^{-1}(a, b, c)$ is contained in the badge region. Then SMT$(A, B, C) \in \Pi^{-1}(a, b, c)$ is also contained in the piping region centered about a tile corresponding to face* 1.

Proof. Assume the setup given in the proof of Proposition 8.1. We will show that the Steiner minimal tree need not meet any of the flaps. By way of contradiction, suppose that SMT(A, B, C) meets the top flap, the flap contained in $T_{(2,1,-1)}$, outside the piping region. If SMT(A, B, C) meets the top flap, then at least one fixed point or vertex of SMT(A, B, C) must lie above m_2. Note that by construction, S is contained in T_0 and cannot be this point. Since the only tile in the badge region which lies above m_2 is a tile corresponding to face 3, the fixed point must lie in the interior of face 3. Thus, A must lie in the top flap outside the piping region. For the remainder of the argument, we will denote A by A_1 and label the other copies of $\Pi^{-1}(a)$, $\Pi^{-1}(b)$, and $\Pi^{-1}(c)$ contained in tiles meeting the badge region as shown in the figure on the top of the next page. We will show either that any Steiner tree SMT(A_1, B_i, C_j) with S in τ_0 contained within the badge region cannot realize SMT(a, b, c) or that there exists another copy of the tree within the piping region.

We will first determine which combinations cannot realize SMT(a, b, c).Once those combinations are determined, we will show that the remaining combinations have an equivalent copy contained in the piping region.

Construct the sector $S_{A_2A_1}$. If any points B_i and C_j are both contained in $S_{A_2A_1}$, they must both be separated from A_1 by $P_{A_2A_1}$. Thus, by Theorem 7.5, we know that $\Pi(\text{SMT}(A_1, B_i, C_j)) \neq \text{SMT}(a, b, c)$ for B_i and C_j contained in these sectors.



By this argument, the combinations (B_i, C_j), for $i = 4, 5$ and $j = 4, 5, 6$, cannot be used with A_1 to realize SMT(a, b, c).

Construct the half-plane $H_{A_3A_1}$. If any points B_i and C_j are both contained in $H_{A_3A_1}$, they must be separated from A_1 by $P_{A_3A_1}$. Thus, by Theorem 7.5, we know that $\Pi(\text{SMT}(A_1, B_i, C_j)) \neq \text{SMT}(a, b, c)$ for B_i and C_j contained in these sectors. By this argument, the combinations (B_i, C_j), for $i = 4, 6$ and $j = 3, 4, 5, 6$, cannot be used with A_1 to realize SMT(a, b, c).

Construct the half-plane $H_{A_4A_1}$. If any points B_i and C_j are both contained in $H_{A_4A_1}$, they must be separated from A_1 by $P_{A_4A_1}$. Thus, by Theorem 7.5, we know that $\Pi(\text{SMT}(A_1, B_i, C_j)) \neq \text{SMT}(a, b, c)$ for B_i and C_j contained in these sectors. By this argument, the combinations (B_i, C_j), for $i = 3, 4, 5, 6$ and $j = 4, 6$, cannot be used with A_1 to realize SMT(a, b, c).

Consider SMT(A_1, B_1, C_3). Note that both A_1 and B_1 must be contained in $S_{C_1C_3}$. Thus, A_1 and B_1 must be separated from C_3 by $P_{C_1C_3}$. By Theorem 7.5, $\Pi(\text{SMT}(A_1, B_1, C_3)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_1, B_1, C_4). Note that both A_1 and B_1 must be contained in $S_{C_2C_4}$. Thus, A_1 and B_1 must be separated from C_4 by $P_{C_2C_4}$. By Theorem 7.5, $\Pi(\text{SMT}(A_1, B_1, C_4)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_1, B_1, C_5). Note that both A_1 and B_1 must be contained in $H_{C_1C_5}$. Thus, A_1 and B_1 must be separated from C_5 by $P_{C_1C_5}$. By Theorem 7.5, $\Pi(\text{SMT}(A_1, B_1, C_5)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_1, B_1, C_6). Note that both A_1 and B_1 must be contained in $S_{C_1C_6}$. Thus, A_1 and B_1 must be separated from C_3 by $P_{C_1C_6}$. By Theorem 7.5, $\Pi(\text{SMT}(A_1, B_1, C_6)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_1, B_2, C_2). Let V be the intersection of m_1 and n_0. Note that V and A_1 are on the same side of $\overleftrightarrow{B_2C_2}$, V and B_2 are on the same side of $\overleftrightarrow{A_1C_2}$, and V and C_2 are on the same side of $\overleftrightarrow{A_1B_2}$. Thus V is contained in $\triangle A_1B_2C_2$. By Theorem 7.4, $\Pi(\text{SMT}(a, b, c)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_1, B_2, C_3). Note that C_3 lies in $H_{A_2A_1}$ and that A_1 lies in $S_{C_1C_3}$. B_2 must lie in at least one of $S_{A_2A_1}$ and $H_{C_1C_3}$. Suppose B_2 lies in $S_{A_2A_1}$. Then both B_2 and C_3 must be separated from A_1 by $P_{A_2A_1}$. If B_2 does not lie in $S_{A_2A_1}$, then B_2 must lie in $H_{C_1C_3}$. But then both B_2 and A_1 must be separated from C_3 by $P_{C_1C_3}$. By Theorem 7.5, $\Pi(\text{SMT}(A_1, B_2, C_3)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_1, B_2, C_4). Note that C_4 lies in $H_{A_3A_1}$ and that A_1 lies in $S_{C_2C_4}$. B_2 must lie in at least one of $H_{A_3A_1}$ and $S_{C_2C_4}$. Suppose B_2 lies in $H_{A_3A_1}$. Then both B_2 and C_4 must be separated from A_1 by $P_{A_3A_1}$. If B_2 does not lie in $H_{A_3A_1}$, then B_2 must lie in $S_{C_2C_4}$. But then both B_2 and A_1 must be separated from C_4 by $P_{C_2C_4}$. By Theorem 7.5, $\Pi(\text{SMT}(A_1, B_2, C_4)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_1, B_2, C_5). Note that C_5 lies in $H_{A_3A_1}$ and that A_1 lies in $H_{C_1C_5}$. B_2 must lie in at least one of $H_{A_3A_1}$ and $H_{C_1C_5}$. Suppose B_2 lies in $H_{A_3A_1}$. Then both B_2 and C_5 must be separated from A_1 by $P_{A_3A_1}$. If B_2 does not lie in $H_{A_3A_1}$, then B_2 must lie in $H_{C_1C_5}$. But then both B_2 and A_1 must be separated from C_5 by $P_{C_1C_5}$. By Theorem 7.5, $\Pi(\text{SMT}(A_1, B_2, C_5)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_1, B_2, C_6). Note that both A_1 and B_2 must be contained in $S_{C_1C_6}$. Thus, A_1 and B_2 must be separated from C_6 by $P_{C_1C_6}$. By Theorem 7.5, $\Pi(\text{SMT}(A_1, B_2, C_6)) \neq \text{SMT}(a, b, c)$.

We now consider the combinations (A_1, B_i, C_j) for $i = 4, 5, 6$ and $j = 1, 2$. By arguments of symmetry, $\Pi(\text{SMT}(A_1, B_i, C_j)) \neq \text{SMT}(a, b, c)$ for $i = 4, 5, 6$ and $j = 1, 2$.

Consider SMT(A_1, B_3, C_3). Let V be the intersection of m_1 and n_0. Note that V and A_1 are on the same side of $\overleftrightarrow{B_3C_3}$, V and B_3 are on the same side of $\overleftrightarrow{A_1C_3}$, and V and C_3 are on the same side of $\overleftrightarrow{A_1B_3}$. Thus V is contained in $\triangle A_1B_3C_3$. By Theorem 7.4, $\Pi(\text{SMT}(A_1, B_1, C_1)) \neq \text{SMT}(a, b, c)$.

The only remaining cases are (A_1, B_1, C_1), (A_1, B_1, C_2), and (A_1, B_2, C_1). We will show that copies of these trees exist within the piping region. However, we will not claim that the Steiner point S must remain in τ_0.

For (A_1, B_1, C_1), note that $\Pi(\text{SMT}(A_1, B_1, C_1)) = \Pi(\text{SMT}(A_2, B_2, C_2))$ since SMT(A_2, B_2, C_2) is a rotation of SMT(A_1, B_1, C_1) about V. SMT(A_2, B_2, C_2) is contained within the piping region.

For (A_1, B_1, C_2), note that $\Pi(\text{SMT}(A_1, B_1, C_2)) = \Pi(\text{SMT}(A_2, B_2, C_1))$ since SMT(A_2, B_2, C_1) is a rotation of SMT(A_1, B_1, C_2) about V. SMT(A_2, B_2, C_1) is contained within the piping region.

For (A_1, B_2, C_1), note that $\Pi(\text{SMT}(A_1, B_2, C_1)) = \Pi(\text{SMT}(A_2, B_1, C_2))$ since SMT(A_2,B_1,C_2) is a rotation of SMT(A_1,B_2,C_1) about V. Also, SMT(A_2,B_1,C_2)

is contained within the piping region.

Thus, each possible combination (A_1, B_i, C_j) does not realize $\mathrm{SMT}(a, b, c)$ or has a copy within the piping region. Likewise, each possible combination involving B_5 or C_5 does not realize $\mathrm{SMT}(a, b, c)$ or has a copy within the piping region. Therefore, there is a solution contained in the piping region. □

The region resulting from Theorem 8.2 is the piping region, which we illustrated in Figure 7.

Reduction to the truncated triangle region. We further subdivide the piping region into the *truncated triangle region* and the *side flaps* (Figure 8).

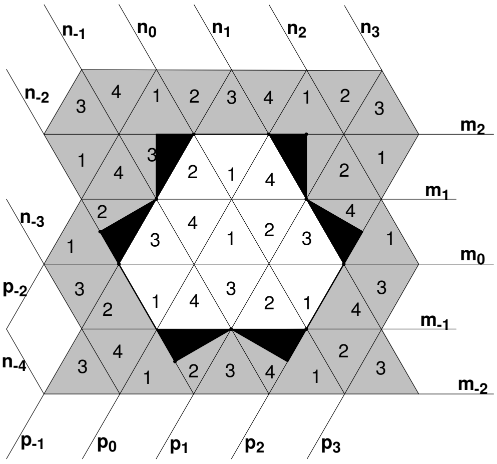
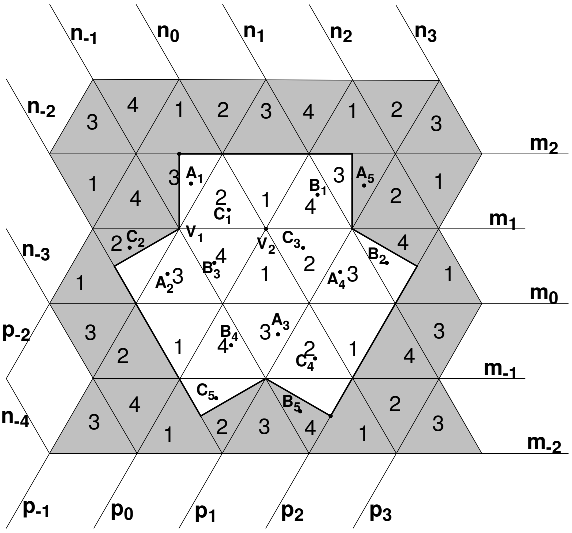


Figure 8. The truncated triangle region (closure of white polygon) and the side flaps (in black).

Theorem 8.3. *Suppose a, b, and c are three points on distinct faces of $\mathcal{T}$, none of which are in the interior of face* 1. *Suppose $\mathrm{SMT}(A, B, C) \in \Pi^{-1}(a, b, c)$ is contained in the piping region. Then either $\mathrm{SMT}(A, B, C) \in \Pi^{-1}(a, b, c)$ is also contained in the truncated triangle region centered about a tile corresponding to face* 1 *or there is a copy of* $\mathrm{SMT}(A, B, C)$ *contained within the truncated triangle region that is a rotation of* $\mathrm{SMT}(A, B, C)$.

Proof. Assume the setup in the proof of Proposition 8.1. Without loss of generality, suppose that $\mathrm{SMT}(A, B, C)$ is in the piping region. We will show that the Steiner minimal tree need not meet any of the side flaps. Although the final cases of the proof of Theorem 8.2 did not guarantee that S was contained in τ_0, S must be contained in the truncated triangle region. This is because all the trees which could be rotated to lie within the piping region contained fixed points contained within the truncated triangle region. Because S must be contained in the convex hull of the triangular region formed from the fixed points, S must be contained within the truncated triangle region.

By way of contradiction, suppose the Steiner minimal tree meets the flap contained in $T_{(2,-1,-1)}$. If $\mathrm{SMT}(A, B, C)$ meets this side flap, then at least one fixed point or vertex of $\mathrm{SMT}(A, B, C)$ must lie above to the left of p_{-1} and above m_1. Since S is contained in the truncated triangle region (Figure 8), S cannot be this point. Since the only tile in the piping region which lies to the left of p_{-1} and above m_1 is a tile corresponding to face 3, the fixed point must lie in the interior of face 3. Thus, A must lie in the specified side flap outside the truncated triangle region. For the remainder of the proof we will denote A by A_1 and number the other points within the piping region as follows:



Construct the sector $S_{A_2A_1}$. If any points B_i and C_j are both contained in $S_{A_2A_1}$, they must both be separated from A_1 by $P_{A_2A_1}$. Thus, by Theorem 7.5, we know that $\Pi(\mathrm{SMT}(A_1, B_i, C_j)) \neq \mathrm{SMT}(a, b, c)$ for B_i and C_j contained in these sectors. By this argument, the combinations (B_i, C_j), for $i = 4, 5$ and $j = 2, 4, 5$, cannot be used with A_1 to realize $\mathrm{SMT}(a, b, c)$.

Construct the half-plane $H_{A_3A_1}$. If any points B_i and C_j are both contained in $H_{A_3A_1}$, they must both be separated from A_1 by $P_{A_3A_1}$. Thus, by Theorem 7.5, we know that $\Pi(\mathrm{SMT}(A_1, B_i, C_j)) \neq \mathrm{SMT}(a, b, c)$ for B_i and C_j contained in these sectors. By this argument, the combinations (B_i, C_j), for $i = 2, 4, 5$ and $j = 4, 5$, cannot be used with A_1 to realize $\mathrm{SMT}(a, b, c)$.

Construct the sector $S_{A_4A_1}$. If any points B_i and C_j are both contained in $S_{A_4A_1}$, they must both be separated from A_1 by $P_{A_4A_1}$. Thus, by Theorem 7.5, we know that $\Pi(\mathrm{SMT}(A_1, B_i, C_j)) \neq \mathrm{SMT}(a, b, c)$ for B_i and C_j contained in these sectors.

By this argument, the combinations (B_i, C_j), for $i = 2, 5$ and $j = 3, 4$, cannot be used with A_1 to realize $\mathrm{SMT}(a, b, c)$.

For $\mathrm{SMT}(A_1, B_1, C_2)$, note that A_1 and B_1 are contained in $S_{C_1C_2}$, so they are both separated from C_2 by $P_{C_1C_2}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_1, C_2)) \neq \mathrm{SMT}(a, b, c)$.

For $\mathrm{SMT}(A_1, B_1, C_4)$, note that A_1 and B_1 are contained in $S_{C_3C_4}$, so they are both separated from C_4 by $P_{C_3C_4}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_1, C_4)) \neq \mathrm{SMT}(a, b, c)$.

For $\mathrm{SMT}(A_1, B_1, C_5)$, note that A_1 and B_1 are contained in $S_{C_1C_5}$, so they are both separated from C_5 by $P_{C_1C_5}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_1, C_5)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_3, C_2)$. Let V_1 be the intersection of m_1 and n_{-1}. Note that A_1 and V_1 are on the same side of $\overleftrightarrow{B_3C_2}$, B_3 and V are on the same side of $\overleftrightarrow{A_1C_2}$, and C_2 and V are on the same side of $\overleftrightarrow{A_1B_3}$. Thus, V_1 is contained in $\triangle ABC$. By Theorem 7.4, $\Pi(\mathrm{SMT}(A_1, B_3, C_2)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_3, C_4)$. Note that A_1 lies in $S_{C_2C_4}$ and C_4 lies in $S_{A_2A_1}$. Note that B_3 must lie in at least one of $S_{C_2C_4}$ and $S_{A_2A_1}$. If B_3 lies in $S_{C_2C_4}$, both B_3 and A_1 must be separated from C_4 by $P_{C_2C_4}$. If B_3 lies in $S_{A_2A_1}$, both B_3 and C_4 must be separated from A_1 by $P_{A_2A_1}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_3, C_4)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_3, C_5)$. Note that both A_1 and B_3 lie in $S_{C_1C_5}$, so they are both separated from C_5 by $P_{C_1C_5}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_3, C_5)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_2, C_1)$. Note that both A_1 and C_1 lie in $S_{B_4B_2}$, so they are both separated from B_2 by $P_{B_4B_2}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_2, C_1)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_2, C_2)$. Note that both A_1 and C_2 lie in $S_{B_4B_2}$, so they are both separated from B_2 by $P_{B_4B_2}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_2, C_2)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_4, C_1)$. Note that both A_1 and C_1 lie in $S_{B_3B_4}$, so they are both separated from B_4 by $P_{B_3B_4}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_4, C_1)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_4, C_3)$. Note that A_1 lies in $S_{B_3B_4}$ and B_4 lies in $H_{A_3A_1}$. Note that C_3 must lie in at least one of $S_{B_3B_4}$ and $H_{A_3A_1}$. If C_3 lies in $S_{B_3B_4}$, both A_1 and C_3 are separated from B_4 by $P_{B_3B_4}$. If C_3 lies in $H_{A_1A_3}$, both B_4 and C_3 are separated from A_1 by $P_{A_3A_1}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_4, C_3)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_5, C_1)$. Note that both A_1 and C_1 lie in $S_{B_5B_4}$, so they are both separated from B_5 by $P_{B_4B_5}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_5, C_1)) \neq \mathrm{SMT}(a, b, c)$.

The only remaining cases are (A_1, B_1, C_1), (A_1, B_1, C_3), and (A_1, B_3, C_3). We will show that copies of these trees exist within the truncated triangle region.

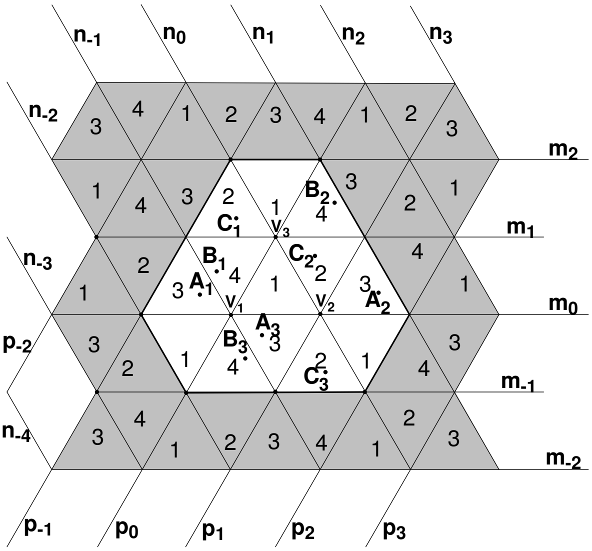
For $\mathrm{SMT}(A_1, B_1, C_1)$, we have $\Pi(\mathrm{SMT}(A_1, B_1, C_1)) = \Pi(\mathrm{SMT}(A_4, B_3, C_3))$ and $\mathrm{SMT}(A_4, B_3, C_3)$ is contained within the truncated triangle region.

For $\mathrm{SMT}(A_1, B_1, C_3)$, we have $\Pi(\mathrm{SMT}(A_1, B_1, C_3)) = \Pi(\mathrm{SMT}(A_4, B_3, C_1))$ and $\mathrm{SMT}(A_4, B_3, C_1)$ is contained within the truncated triangle region.

For (A_1, B_3, C_3), we have $\Pi(\mathrm{SMT}(A_1, B_3, C_3)) = \Pi(\mathrm{SMT}(A_4, B_1, C_1))$ and $\mathrm{SMT}(A_4, B_1, C_1)$ is contained within the truncated triangle region.

Thus, each possible combination (A_1, B_i, C_j) does not realize $\mathrm{SMT}(a, b, c)$ or has a copy within the truncated triangle region. Likewise, each possible combination involving A_5, B_2, B_5, C_5, or C_2 cannot realize $\mathrm{SMT}(a, b, c)$ or has a copy within the truncated triangle region. Therefore, there is a solution contained in the truncated triangle region. □

Final reductions. Within the truncated triangle region, there are three copies of every face that contains a terminal point (the center of each region does not contain any points; in this scenario, face 1). That means that there are three copies of each point:



If all combinations of three points were considered possible configurations for the Steiner minimal tree, there would be 27 different Steiner trees that could be considered. However, some of these possibilities may still be eliminated.

There are three remaining combinations that can be eliminated within the truncated triangle region. Let V_1 be the intersection of m_0 and n_{-2}, V_2 be the intersection of m_0 and n_0, and V_3 be the intersection of m_1 and n_0.

Consider $\mathrm{SMT}(A_1, B_3, C_2)$. Since A_1 and V_1 are on the same side of $\overleftrightarrow{B_3C_2}$, B_3 and V_1 are on the same side of $\overleftrightarrow{A_1C_2}$, and C_2 and V_1 are on the same $\overleftrightarrow{A_1B_3}$, then V_1 is contained in the interior of $\triangle A_1B_3C_2$. By Theorem 7.4, $\Pi(\mathrm{SMT}(A_1, B_3, C_2)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_2, B_1, C_3)$. Since A_2 and V_2 are on the same side of $\overleftrightarrow{B_1C_3}$, B_1 and V_2 are on the same side of $\overleftrightarrow{A_2C_3}$, and C_3 and V_2 are on the same $\overleftrightarrow{A_2B_1}$, then V_2 is contained in the interior of $\triangle A_2B_1C_3$. By Theorem 7.4, $\Pi(\mathrm{SMT}(A_2, B_1, C_3)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_3, B_2, C_1)$. Since A_3 and V_3 are on the same side of $\overleftrightarrow{B_2C_1}$, B_2 and V_3 are on the same side of $\overleftrightarrow{A_3C_1}$, and C_1 and V_3 are on the same $\overleftrightarrow{A_3B_2}$, then V_3 is contained in the interior of $\triangle A_3B_2C_1$. By Theorem 7.4, $\Pi(\mathrm{SMT}(A_3, B_2, C_1)) \neq \mathrm{SMT}(a, b, c)$.

List of potential combinations in case 2. The remaining possibilities are

$$\begin{array}{lll}
(A_1, B_1, C_1), & (A_2, B_1, C_1), & (A_3, B_1, C_1), \\
(A_1, B_1, C_2), & (A_2, B_1, C_2), & (A_3, B_1, C_2), \\
(A_1, B_1, C_3), & (A_2, B_2, C_1), & (A_3, B_1, C_3), \\
(A_1, B_2, C_1), & (A_2, B_2, C_2), & (A_3, B_2, C_2), \\
(A_1, B_2, C_2), & (A_2, B_2, C_3), & (A_3, B_2, C_3), \\
(A_1, B_2, C_3), & (A_2, B_3, C_1), & (A_3, B_3, C_1), \\
(A_1, B_3, C_1), & (A_2, B_3, C_2), & (A_3, B_3, C_2), \\
(A_1, B_3, C_3), & (A_2, B_3, C_3), & (A_3, B_3, C_3).
\end{array}$$

Thus, the Steiner tree which realizes $\mathrm{SMT}(a, b, c)$ will be formed from one of the 24 combinations in this list.

9. Case 3: Three points on two faces

In this section, we consider the cases that haven't been addressed in the other sections, namely where three points lie on two faces and cannot be considered to lie on three faces or a single face. The two remaining possibilities are:

(1) Two of the points are contained in the interior of one face with the third point anywhere not meeting that face.

(2) One point is contained in the interior of a face f, a second point is contained in the interior of an edge adjacent to f, and the final point is in the complement of f.

The arguments for both are the same.

We will assume a and b are on the same face and that at least a is in the interior of the face. Thus either b is in the interior of the face or in the interior of an edge of the face.

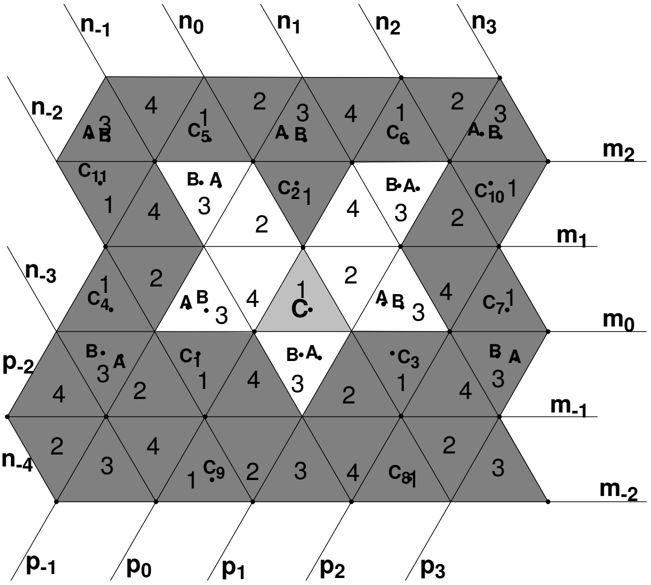
On the tiled plane, there are infinitely many copies of $A \in \Pi^{-1}(a)$ and $B \in \Pi^{-1}(b)$. Suppose $\mathrm{SMT}(A, B, C)$ realizes $\mathrm{SMT}(a, b, c)$. Then either A and B reside on the same tile, or they don't. We will discuss each case separately. We will discuss the former case here and the latter starting on page 392.

A and B on the same tile. In this case, the following theorem provides a region containing the fixed points that can realize $\mathrm{SMT}(a, b, c)$:

Theorem 9.1. *Let $a, b, c \in \mathcal{T}$ and assume*

$$A \in \Pi^{-1}(a), \quad B \in \Pi^{-1}(b), \quad C \in \Pi^{-1}(c)$$

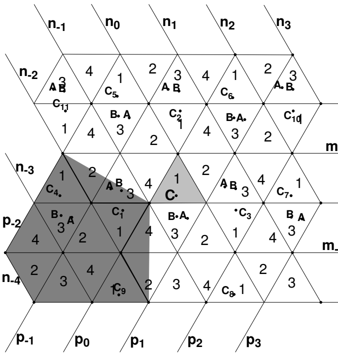
are the points that determine $\mathrm{SMT}(a, b, c)$. *If A and B are on the same tile, the Steiner minimal tree must be contained in the ten-triangle region shown here in white and light gray*:



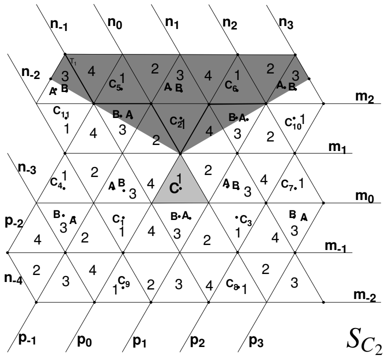
Proof. We can assume without loss of generality that suppose c is on face 1, while a and b are on face 3. We suppose that C is contained in the light gray tile in the figure above.

Case 1: Suppose C is not on a vertex of a tile. The other copies of $C_i \in \Pi^{-1}(c)$ are located on the other tiles corresponding to face 1. We number them as in the figure above. We will now determine the tiles on which A and B could possibly reside.

Construct S_{C_1C}. The points A_i and B_j which lie in S_{C_1C} must be separated from C by P_{C_1C}. By Theorem 7.5, $\mathrm{SMT}(A_i, B_j, C)$ cannot realize $\mathrm{SMT}(a, b, c)$. Thus, we can eliminate from consideration as a candidate for containing A and B any tiles whose interior overlaps the region S_{C_1C}, which we show in dark gray (left):



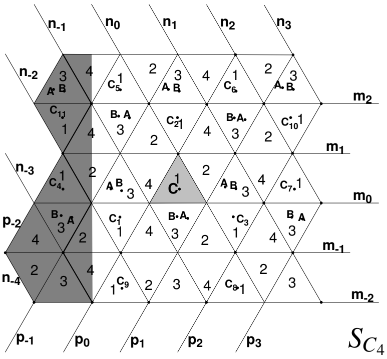
S_{C_1C}



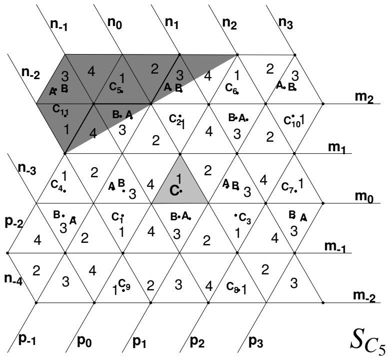
S_{C_2C}

Construct S_{C_2C}. Again, using Theorem 7.5, we can eliminate any tile contained in S_{C_2C}, which is the reason shown in dark gray in the figure above and to the right.

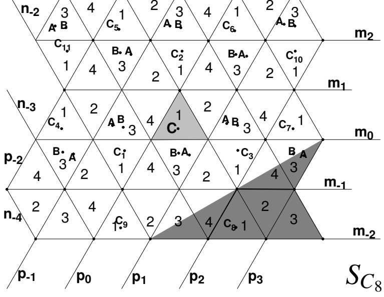
Continue the process by constructing the sectors S_{C_iC}, where $i = 3, \ldots, 9$. Three of these are shown below, while the other four are obtainable by reflection in a vertical line (through the central triangle) from others already illustrated: S_{C_3C} from S_{C_1C}, S_{C_6C} from S_{C_5C}, S_{C_7C} from S_{C_4C}, and S_{C_9C} from S_{C_8C}.



S_{C_4C}



S_{C_5C}



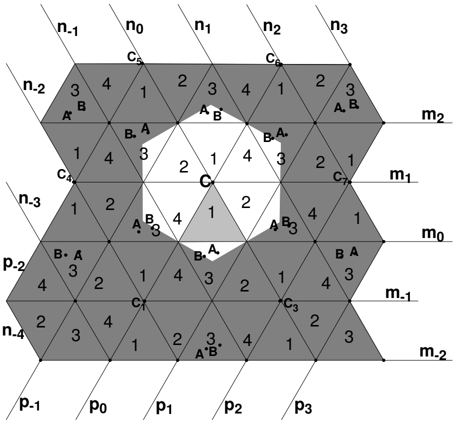
S_{C_8C}

The only copies of tile 3 not completely covered by the union of the shaded regions are those contained in the white region in the statement of Theorem 9.1.

By hypothesis, A and B are contained on the same tile. The convex hull of $\triangle ABC$ contains the tree realizing $\mathrm{SMT}(a, b, c)$. The white region is the minimal collection of tiles containing all such possible convex hulls. Since there are five tiles containing copies of A and B in this region, there are five potential Steiner trees which must be tested within this region.

Case 2: Suppose C is a vertex of a tile. It can only be the vertex at the intersection of m_1 and n_0, because the other vertices are adjacent to tiles containing A and B, and this case has already been addressed in Section 6.

Construct $\tilde{H}_{C_iC}$ for $i = 1, 3, 4, 5, 6, 7$. The union of the added "union of the" shaded regions $\tilde{H}_{C_iC}$ is shown here:



If both A_j and B_k lie in any $\tilde{H}_{C_iC}$, they must both be separated from C by P_{C_iC}. By Theorem 7.5, A_j and B_j cannot be used with C to realize $\mathrm{SMT}(a, b, c)$. Note that at least one of A and B must lie in the unshaded region, and A and B are on the same tile by hypothesis. Thus, there are six possible path networks that connect C with a pair of points A_j and B_k which are contained on the same tile where at least one is not in the shaded region. Since each path has one identical path by reflection across C, there are only three distinct paths, and there exists a copy of each in the region stated in Theorem 9.1. □

A and B not on the same tile. We now study the case that A and B are not on the same tile. This will occupy us through page 399. We will determine the faces where the Steiner point can reside in Theorem 9.2. We will then find the region that must contains the fixed points. We will eliminate possibilities for fixed points in Theorems 9.3–9.6. We will then make final reductions and list the combinations that could realize $\mathrm{SMT}(a, b, c)$.

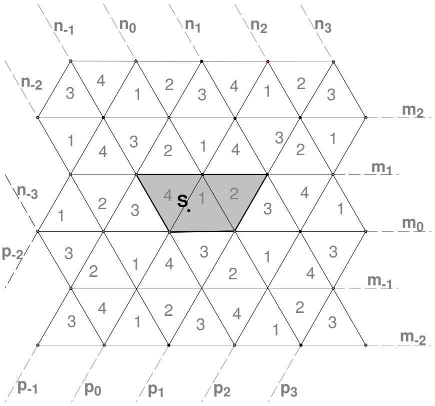
Theorem 9.2. *Assume the setup in Theorem 9.1. Suppose that s is the Steiner point for* $\mathrm{SMT}(a, b, c)$. *If A and B are not found on the same tile, then s can not be on the face containing a and b, including the interior of its edges.*

Proof. By way of contradiction, suppose s is on the same face as a and b. Suppose $S \in \Pi^{-1}(s)$ is contained in the region bounded by n_{-1}, m_1, and p_1. Without loss of generality, c is on face 4, and a and b are on face 3.

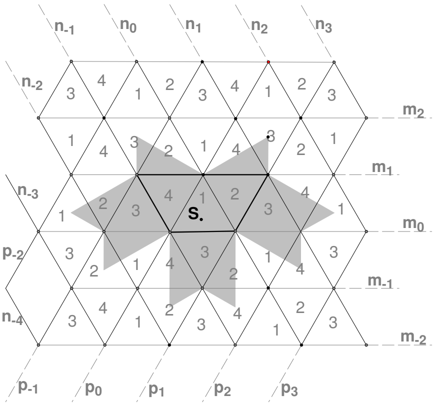
Case 1: Suppose S is not on the same tile as B. Then there exists a distinct point $S' \in \Pi^{-1}(s)$ on the same tile as B. By Lemma 7.1, $S'B < SB$. Then $P_{SS'}$ separates B from S. By Theorem 7.8, S cannot be the Steiner point.

Case 2: Suppose S is on the same tile as B, but S is not a vertex. Then there exists an $S' \in \Pi^{-1}(s)$ on the same tile as A. Since S is not a vertex, $S' \neq S$. Then $P_{SS'}$ separates A from S. By Theorem 7.8, S cannot be the Steiner point. □

It follows from Theorem 9.2 that s must be contained on at least one of faces 1, 2, or 4. Since S cannot be on any tile corresponding to face 3, we can fix S in the shaded region bounded by m_1, m_0, n_{-1}, and p_1, which we call the *key trapezoid*:

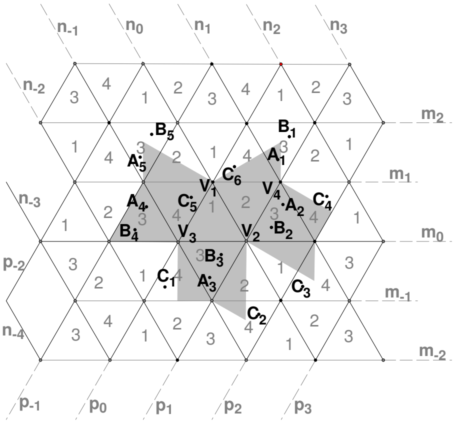


By a similar procedure to that discussed on pages 378 and following, we can eliminate all points lying in the sectors $S_{S'S}$ or half-planes $H_{S'S}$ for all $S' \neq S$, where $S' \in \Pi^{-1}(s)$. The resulting region is this:



Because no terminals are located on faces 1 or 2, the Steiner tree will never cross the copies of face 1 or 2 whose interior meets the edge of this region. We can eliminate these to obtain the region shaded in the figure below. Within this region, there are a maximum of four copies of A, four copies of B, and five copies of C, resulting in a maximum of 80 possible Steiner trees. However, we can reduce the region even further.

Theorem 9.3. *Suppose S is contained in the key trapezoid* (page 393). *Let C_1 and A_i, B_j (with $i, j = 1, \dots, 5$) lie in the triangles specified in this diagram*:



Then $\Pi(\mathrm{SMT}(A_i, B_j, C_1)) \neq \mathrm{SMT}(a, b, c)$ *for all i, j with $i \neq j$. Hence, the tile containing C_1 can be removed from the region of interest.*

Proof. The last assertion follows immediately once we show that no combination (A_i, B_j, C_1) which can be used to realize $\mathrm{SMT}(a, b, c)$. We analyze each case:

Consider $\mathrm{SMT}(A_1, B_2, C_1)$. Both B_2 and C_1 are contained in $S_{A_2A_1}$. Thus B_2 and C_1 are separated from A_1 by $P_{A_2A_1}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_2, C_1)) \neq \mathrm{SMT}(a, b, c)$. Similarly, $\Pi(\mathrm{SMT}(A_2, B_1, C_1)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_3, C_1)$. Both B_3 and C_1 are contained in $H_{A_3A_1}$. Thus B_3 and C_1 are separated from A_1 by $P_{A_3A_1}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_3, C_1)) \neq \mathrm{SMT}(a, b, c)$. Similarly, $\Pi(\mathrm{SMT}(A_3, B_1, C_1)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_4, C_1)$. Both B_4 and C_1 are contained in $H_{A_4A_1}$. Thus B_4 and C_1 are separated from A_1 by $P_{A_4A_1}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_4, C_1)) \neq \mathrm{SMT}(a, b, c)$. Similarly, $\Pi(\mathrm{SMT}(A_4, B_1, C_1)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_5, C_1)$. Both A_1 and B_5 are contained in $S_{C_5C_1}$. Thus A_1 and B_5 are separated from C_1 by $P_{C_5C_1}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_5, C_1)) \neq \mathrm{SMT}(a, b, c)$. Similarly, $\Pi(\mathrm{SMT}(A_5, B_1, C_1)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_2, B_3, C_1)$. Both B_3 and C_1 are contained in $S_{A_3A_2}$. Thus B_3 and C_1 are separated from A_1 by $P_{A_3A_2}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_2, B_3, C_1)) \neq \mathrm{SMT}(a, b, c)$. Similarly, $\Pi(\mathrm{SMT}(A_3, B_2, C_1)) \neq \mathrm{SMT}(a, b, c)$.

Consider SMT(A_2, B_4, C_1). We claim that V_3 is contained in the interior of $\triangle A_2B_4C_1$. Note that V_3 and C_1 are on the same side of $\overleftrightarrow{A_2B_4}$, V_3 and B_4 are on the same side of $\overleftrightarrow{A_2C_1}$, and V_3 and A_2 are on the same side of $\overleftrightarrow{A_2B_4}$. Thus $\triangle A_2B_4C_1$ contains V_3. By Theorem 7.4, $\Pi(\text{SMT}(A_2, B_4, C_1)) \neq \text{SMT}(a, b, c)$. Similarly, $\Pi(\text{SMT}(A_4, B_2, C_1)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_2, B_5, C_1). Both A_2 and C_1 are contained in $H_{B_3B_5}$. Thus A_2 and C_1 are separated from B_5 by $P_{B_3B_5}$. By Theorem 7.5, $\Pi(\text{SMT}(A_2, B_5, C_1)) \neq \text{SMT}(a, b, c)$. Similarly, $\Pi(\text{SMT}(A_5, B_2, C_1)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_3, B_4, C_1). Both A_3 and C_1 are contained in $S_{B_3B_4}$. Thus A_3 and C_1 are separated from B_4 by $P_{B_3B_4}$. By Theorem 7.5, $\Pi(\text{SMT}(A_3, B_4, C_1)) \neq \text{SMT}(a, b, c)$. Similarly, $\Pi(\text{SMT}(A_3, B_4, C_1)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_3, B_5, C_1). Both A_3 and C_1 are contained in $H_{B_3B_5}$. Thus A_3 and C_1 are separated from B_5 by $P_{B_3B_5}$. By Theorem 7.5, $\Pi(\text{SMT}(A_3, B_5, C_1)) \neq \text{SMT}(a, b, c)$. Similarly, $\Pi(\text{SMT}(A_5, B_2, C_1)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_4, B_5, C_1). Both A_4 and C_1 are contained in $S_{B_4B_5}$. Thus A_4 and C_1 are separated from B_5 by $P_{B_4B_5}$. By Theorem 7.5, $\Pi(\text{SMT}(A_4, B_5, C_1)) \neq \text{SMT}(a, b, c)$. Similarly, $\Pi(\text{SMT}(A_5, B_4, C_1)) \neq \text{SMT}(a, b, c)$. □

Theorem 9.4. *Suppose S is contained in the key trapezoid* (page 393). *Let C_2 and A_i, B_j (with $i, j = 1, \dots, 5$) lie in the triangles specified in the diagram of Theorem 9.3. Then $\Pi(\text{SMT}(A_i, B_j, C_2)) \neq \text{SMT}(a, b, c)$ for all i, j with $i \neq j$. That is, the tile containing C_2 can be removed from the region of interest.*

Proof. Again we apply a case-by-case analysis.

Consider SMT(A_1, B_2, C_2). Both B_2 and C_2 are contained in $S_{A_2A_1}$. Thus, B_2 and C_2 are separated from A_1 by $P_{A_1A_2}$. By Theorem 7.5, $\Pi(\text{SMT}(A_1, B_2, C_2)) \neq (\text{SMT}(a, b, c))$. By a similar argument, $\Pi(\text{SMT}(A_2, B_1, C_2)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_1, B_3, C_2). Both B_3 and C_2 are contained in $H_{A_3A_1}$. Thus, B_3 and C_2 are separated from A_1 by $P_{A_1A_3}$. By Theorem 7.5, $\Pi(\text{SMT}(A_1, B_3, C_2)) \neq \text{SMT}(a, b, c)$. By a similar argument, $\Pi(\text{SMT}(A_3, B_1, C_2)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_1, B_4, C_2). Both B_4 and C_2 are contained in $H_{A_3A_1}$. Thus, B_4 and C_2 are separated from A_1 by $P_{A_1A_3}$. By Theorem 7.5, $\Pi(\text{SMT}(A_1, B_4, C_2)) \neq \text{SMT}(a, b, c)$. By a similar argument, $\Pi(\text{SMT}(A_4, B_1, C_2)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_1, B_5, C_2). We claim that V_1 is contained in the interior of $\triangle A_1B_5C_2$. Both C_2 and V_1 lie on the same side of $\overleftrightarrow{A_1B_5}$, B_5 and V_1 lie on the same side of $\overleftrightarrow{A_1C_2}$, and A_1 and V_1 lie on the same side of $\overleftrightarrow{B_5C_2}$. Thus V_1 must be contained in the interior of $\triangle A_1B_5C_2$. By Theorem 7.4, $\Pi(\text{SMT}(A_1, B_5, C_2)) \neq \text{SMT}(a, b, c)$. By a similar argument, $\Pi(\text{SMT}(A_5, B_1, C_2)) \neq \text{SMT}(a, b, c)$.

Consider SMT(A_2, B_3, C_2). Recall that S are contained in the convex hull of $\triangle A_2B_3C_2$. By hypothesis, S is contained in the key trapezoid (page 393). These two conditions are satisfied only if $\overline{A_2B_3}$ lies above the vertex V_2. Thus, C_2 and

V_2 lie on the same side of $\overleftrightarrow{A_2B_3}$, B_3 and V_2 lie on the same side of $\overleftrightarrow{A_2C_2}$, and A_2 and V_2 lie on the same side of $\overleftrightarrow{B_3C_2}$. Thus V_2 are contained in the interior of $\triangle A_2B_3C_2$. By Theorem 7.4, $\Pi(\mathrm{SMT}(A_2, B_3, C_2)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_3, B_2, C_2)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_2, B_4, C_2)$. Both A_2 and C_2 are contained in $S_{B_3B_4}$. Thus, A_2 and C_2 are separated from B_4 by $P_{B_3B_4}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_2, B_4, C_2)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_4, B_2, C_2)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_2, B_5, C_2)$. Both A_2 and C_2 are contained in $H_{B_3B_4}$. Thus, A_2 and C_2 are separated from B_5 by $P_{B_2B_5}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_2, B_5, C_2)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_5, B_2, C_2)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_3, B_4, C_2)$. Both A_3 and C_2 are contained in $S_{B_3B_4}$. Thus, A_3 and C_2 are separated from B_4 by $P_{B_3B_4}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_3, B_4, C_2)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_4, B_3, C_2)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_3, B_5, C_2)$. Both A_3 and C_2 are contained in $H_{B_3B_5}$. Thus, A_3 and C_2 are separated from B_5 by $P_{B_3B_5}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_3, B_5, C_2)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument,$\Pi(\mathrm{SMT}(A_5, B_3, C_2)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_4, B_5, C_2)$. Both A_4 and C_2 are contained in $S_{B_4B_5}$. Thus, A_4 and C_2 are separated from B_5 by $P_{B_4B_5}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_4, B_5, C_2)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument,$\Pi(\mathrm{SMT}(A_5, B_4, C_2)) \neq \mathrm{SMT}(a, b, c)$. □

Theorem 9.5. *Suppose S is contained in the key trapezoid* (page 393). *Let C_4 and A_i, B_j (with $i, j = 1, \ldots, 5$) lie in the triangles specified in the diagram of Theorem 9.3. Then $\Pi(\mathrm{SMT}(A_i, B_j, C_4)) \neq \mathrm{SMT}(a, b, c)$ for all i, j with $i \neq j$. That is, the tile containing C_4 can be removed from the region of interest.*

Proof. Consider $\mathrm{SMT}(A_1, B_2, C_4)$. Both B_2 and C_4 are contained in $S_{A_2A_1}$, so B_2 and C_4 are separated from A_1 by $P_{A_1A_2}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_2, C_4)) \neq \mathrm{SMT}(a, b, c))$. By a similar argument, $\Pi(\mathrm{SMT}(A_2, B_1, C_4)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_3, C_4)$. Let V_4 be the intersection of m_1 and p_1. Note that C_4 and V_4 lie on the same side of $\overleftrightarrow{A_1B_3}$, B_3 and V_4 lie on the same side of $\overleftrightarrow{A_1C_4}$, and A_1 and V_4 lie on the same side of $\overleftrightarrow{B_3C_4}$. Thus V_4 are contained in the interior of $\triangle A_1B_3C_4$. By Theorem 7.4, $\Pi(\mathrm{SMT}(A_1, B_3, C_4)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_3, B_1, C_4)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_4, C_4)$. Both A_1 and C_4 are contained in $H_{B_1B_4}$. Thus, A_1 and C_4 are separated from B_4 by $P_{B_1B_4}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_4, C_4)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_4, B_1, C_4)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_5, C_4)$. Both A_1 and C_4 are contained in $H_{B_1B_5}$. Thus, A_1 and C_4 are separated from B_5 by $P_{B_1B_5}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_5, C_4)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_5, B_1, C_4)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_2, B_3, C_4)$. Both A_2 and C_4 are contained in $S_{B_2B_3}$. Thus, A_2

and C_4 are separated from B_3 by $P_{B_2B_3}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_2, B_3, C_4)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_3, B_2, C_4)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_2, B_4, C_4)$. Both A_2 and C_4 are contained in $H_{B_2B_4}$. Thus, A_2 and C_4 are separated from B_4 by $P_{B_2B_4}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_2, B_4, C_4)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_4, B_2, C_4)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_2, B_5, C_4)$. Both A_2 and C_4 are contained in $H_{B_2B_5}$. Thus, A_2 and C_4 are separated from B_5 by $P_{B_2B_5}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_2, B_5, C_4)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_5, B_2, C_4)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_3, B_4, C_4)$. Both A_3 and C_4 are contained in $S_{B_3B_4}$. Thus, A_3 and C_4 are separated from B_4 by $P_{B_3B_4}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_3, B_4, C_4)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_4, B_3, C_4)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_3, B_5, C_4)$. Both A_3 and C_4 are contained in $H_{B_3B_5}$. Thus, A_3 and C_4 are separated from B_5 by $P_{B_3B_5}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_3, B_5, C_4)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_5, B_3, C_4)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_4, B_5, C_4)$. Both A_4 and B_5 are contained in $H_{C_5C_4}$. Thus, A_4 and B_5 are separated from C_4 by $P_{C_5C_4}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_4, B_5, C_4)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_5, B_4, C_4)) \neq \mathrm{SMT}(a, b, c)$. □

Theorem 9.6. *Suppose S is contained in the key trapezoid* (page 393). *Let C_3 and A_i, B_j (with $i, j = 1, \dots, 5$) lie in the triangles specified in the diagram of Theorem 9.3. Then $\Pi(\mathrm{SMT}(A_i, B_j, C_3)) \neq \mathrm{SMT}(a, b, c)$ for all i, j with $i \neq j$. That is, the tile containing C_4 can be removed from the region of interest.*

Proof. Consider $\mathrm{SMT}(A_1, B_2, C_3)$. Both B_2 and C_3 are contained in $S_{A_2A_1}$, so B_2 and C_3 are separated from A_1 by $P_{A_1A_2}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_2, C_3)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_2, B_1, C_3)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_3, C_3)$. Assume B_3 is not a vertex. We claim that V_2 is contained in $\triangle A_1B_3C_2$. Let V_2 be the intersection of m_0 and p_1. Note that V_2 and C_3 are on the same side of $\overleftrightarrow{A_1B_3}$, V_2 and B_3 are on the same side of $\overleftrightarrow{A_1C_3}$, and V_2 and A_1 are on the same side of $\overleftrightarrow{B_3C_3}$. Thus $\triangle A_1B_3C_3$ must contain V_2. By Theorem 7.4, $\Pi(\mathrm{SMT}(A_1, B_3, C_3)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_3, B_1, C_3)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_4, C_3)$. Both A_1 and B_4 are contained in $H_{C_6C_3}$. Thus, A_1 and B_4 are separated from C_3 by $P_{C_6C_3}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_4, C_3)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_4, B_1, C_3)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_5, C_3)$. Both A_1 and C_3 are contained in $H_{B_1B_5}$. Thus, A_1 and C_3 are separated from B_5 by $P_{B_1B_5}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_5, C_3)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_1, B_5, C_3)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_2, B_3, C_3)$. Recall that S are contained in the convex hull of $\triangle A_2B_3C_3$. By hypothesis, S is contained in the key trapezoid (page 393). These two conditions are satisfied only if $\overline{A_2B_3}$ lies above the vertex V_2, the intersection of m_0 and p_1. Thus, C_3 and V_2 lie on the same side of $\overleftrightarrow{A_2B_3}$, B_3 and V_2 lie on the same side of $\overleftrightarrow{A_2C_3}$, and A_2 and V_2 lie on the same side of $\overleftrightarrow{B_3C_3}$. Thus V_2 are contained in the interior of $\triangle A_2B_3C_3$. By Theorem 7.4, $\Pi(\mathrm{SMT}(A_2, B_3, C_3)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_3, B_2, C_3)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_2, B_4, C_3)$. Both A_2 and C_3 are contained in $H_{B_2B_4}$. Thus, A_2 and C_3 are separated from B_4 by $P_{B_2B_4}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_2, B_4, C_3)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_4, B_2, C_3)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_2, B_5, C_3)$. Both A_2 and C_3 are contained in $H_{B_2B_5}$. Thus A_2 and C_3 are separated from B_5 by $P_{B_2B_5}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_2, B_5, C_3)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_5, B_2, C_3)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_3, B_4, C_3)$. Both A_3 and C_3 are contained in $S_{B_3B_4}$. thus, A_3 and C_3 are separated from B_4 by $P_{B_3B_4}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_3, B_4, C_3)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_4, B_3, C_3)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_3, B_5, C_1)$. Both A_3 and C_3 are contained in $H_{B_3B_5}$. Thus, A_3 and C_3 are separated from B_5 by $P_{B_3B_5}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_3, B_5, C_3)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_5, B_3, C_3)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_4, B_5, C_3)$. Both A_4 and B_5 are contained in $H_{C_1C_3}$. Thus A_4 and B_5 are separated from C_3 by $P_{C_1C_3}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_4, B_5, C_3)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_4, B_3, C_3)) \neq \mathrm{SMT}(a, b, c)$. □

The final region of interest, after the removal of the tiles containing C_1, C_2, C_3 and C_4, is shown in Figure 9. This region also contains each of the five Steiner trees that can be considered when A and B are on the same tile (see Theorem 9.1).

Final reductions. The region shown in Figure 9 must contain at least one copy of the tree $\mathrm{SMT}(A, B, C)$ that realizes $\mathrm{SMT}(a, b, c)$ where A and B come from different tiles. Within this region, there are still combinations that can never realize

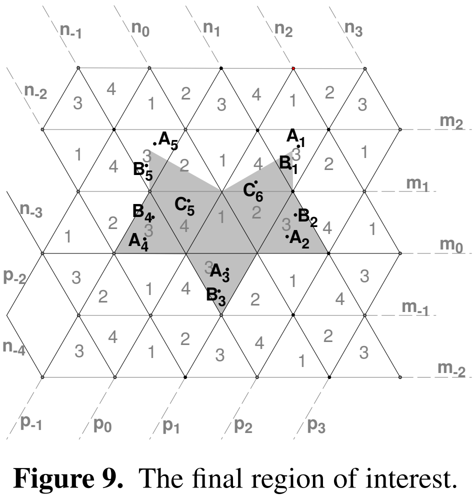


Figure 9. The final region of interest.

$\mathrm{SMT}(a, b, c)$ and thus do not need to be considered. In this section we will eliminate these combinations and then provide a list of all the trees $\mathrm{SMT}(A_i, B_j, C_k)$ that must be considered to determine the $\mathrm{SMT}(A, B, C)$ realizing $\mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_5, C_6)$. Both B_5 and C_6 lie in $H_{A_1A_5}$. Thus, B_5 and C_6 must be separated from A_5 by $P_{A_1A_5}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_5, C_6)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_5, B_1, C_6)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_2, B_5, C_6)$. Both A_2 and C_6 lie in $S_{B_2B_5}$. Thus, A_2 and C_6 must be separated from B_5 by $P_{B_2B_5}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_2, B_5, C_6)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_5, B_2, C_6)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_3, B_5, C_6)$, and let $V_1 = m_1 \cap n_0$. Then A_3 and V_1 are on the same side of $\overleftrightarrow{B_5C_6}$, B_5 and V_1 on the same side of $\overleftrightarrow{A_3C_6}$, and C_6 and V_1 on the same side of $\overleftrightarrow{A_3B_5}$. Thus, $V_1 \subset \triangle A_3B_5C_6$. By Theorem 7.4, $\Pi(\mathrm{SMT}(A_3, B_5, C_6)) \neq \mathrm{SMT}(a, b, c)$. Similarly, $\Pi(\mathrm{SMT}(A_5, B_3, C_6)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_4, B_5, C_6)$. Note that if B_5 is within the shaded region (which is required for it to even be considered), then both B_5 and A_4 lie in $S_{C_5C_6}$. Thus, B_5 and A_4 are separated from C_6 by $P_{C_5C_6}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_4, B_5, C_6)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_5, B_4, C_6)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_4, C_5)$. Both B_4 and C_5 lie in $S_{A_4A_1}$. Thus, both B_4 and C_5 must be separated from A_1 by $P_{A_4A_1}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_4, C_5)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_4, B_1, C_6)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_2, B_5, C_5)$. Both B_5 and C_5 lie in $S_{A_5A_2}$. Thus, both B_5 and C_5 must be separated from A_2 by $P_{A_5A_2}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_2, B_5, C_5)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_5, B_2, C_6)) \neq \mathrm{SMT}(a, b, c)$.

Consider $\mathrm{SMT}(A_1, B_5, C_5)$. Note that if B_5 is within the shaded region, then both B_5 and C_5 are contained in $S_{A_4A_1}$. Thus, B_5 and C_5 are separated from A_1 by $P_{A_4A_1}$. By Theorem 7.5, $\Pi(\mathrm{SMT}(A_1, B_5, C_5)) \neq \mathrm{SMT}(a, b, c)$. By a similar argument, $\Pi(\mathrm{SMT}(A_5, B_1, C_6)) \neq \mathrm{SMT}(a, b, c)$.

List of potential combinations in Case 3. The remaining combinations (A_i, B_j, C_k) for both A and B on the same tile and A and B not on the same tile are

$(A_1, B_2, C_6) \cong (A_4, B_5, C_5)$, $(A_2, B_1, C_6) \cong (A_5, B_4, C_5)$, (A_1, B_3, C_6),
(A_3, B_1, C_6), (A_1, B_4, C_6), (A_4, B_1, C_6),
(A_2, B_3, C_6), (A_3, B_2, C_6), (A_2, B_4, C_6),
(A_4, B_2, C_6), (A_3, B_4, C_6), (A_4, B_3, C_6),
(A_2, B_1, C_5), (A_1, B_2, C_5), (A_1, B_3, C_5),
(A_3, B_1, C_5), (A_2, B_3, C_5), (A_3, B_2, C_5),
(A_2, B_4, C_5), (A_4, B_2, C_5), (A_3, B_4, C_5),
(A_4, B_3, C_5), (A_3, B_5, C_5), (A_5, B_3, C_5),
(A_2, B_2, C_5), (A_3, B_3, C_5), $(A_4, B_4, C_5) \cong (A_1, B_1, C_6)$,
$(A_5, B_5, C_5) \cong (A_2, B_2, C_6)$, (A_3, B_3, C_6).

Thus, the Steiner tree which realizes $\mathrm{SMT}(a, b, c)$ will be formed from one of the 29 combinations included in this list.

10. An algorithm for finding a shortest network on three points

At the end of Sections 8 and 9 we provided lists of combinations which could realize $\mathrm{SMT}(a, b, c)$ for the different cases. In this section we discuss how these lists can be further reduced by considerations of the specific positioning of the points within the faces. We provide two principles upon which the reductions are based. We also provide an algorithm that uses these principles. When the algorithm is applied, we have found that most point combinations can be eliminated.

Two principles allow us to eliminate potential combinations of points from consideration:

- We demonstrated that for Case 2 a solution must reside in the truncated triangle region (Figure 8) and for Case 3 it must resided in the shaded region in Figure 9. In either case, if a point lies outside the corresponding region, no combinations involving that particular point need to be considered.
- If any two points of a combination are separated from the third point by the perpendicular bisector of the third point and a rotation and/or translation of the third point, that combination does not need to be considered (see Theorem 7.5). Recall from Definition 4.2 that for any points P and Q, $\tilde{H}_{PQ} = \{X \mid PX \leq QX\}$. Thus, equivalently, if A and B are contained in $\tilde{H}_{C'C}$ for some $C, C' \in \Pi^{-1}(c)$, then (A, B, C) does not need to be considered.

Using these principles, point combinations within the list can be eliminated from consideration. A systematic approach to the elimination is introduced in the following algorithm.

Algorithm 10.1. The following algorithm provides a shortest network connecting three given points on a regular tetrahedron $\mathcal{T}$.

(1) Determine whether Case 1, 2, or 3 applies.

<u>*Case 1:*</u> If all three points can be considered to lie on a common face, the Steiner tree is just a shortest network on that face (Section 6), and the Steiner tree can be constructed using Algorithm 2.1. The algorithm is complete.

<u>*Case 2:*</u> If the three points can be considered to lie on distinct faces of $\mathcal{T}$, define the region of interest to be the truncated triangle region (Figure 8). Define the list of potential combinations to be the list on page 389. Label the faces so that the face not considered to contain any points is face 1. Proceed to Steps (2)–(4).

Case 3: Otherwise, define the shaded region to be that shown in Figure 9. Define the list of potential combinations to be the list on page 399. Label the faces so that the face considered to contain two points is face 3, and the face considered to contain one point is face 4. Proceed to Steps (2)–(4).

(2) Eliminate any combinations within the list of potential combinations that contain points which are not contained within the shaded region.

(3) For all C_m contained in the shaded region:

(a) For all $C_i \neq C_m$ in the shaded region, construct $\tilde{H}_{C_i C_m}$. Eliminate any combinations (A_k, B_l, C_m) where A_k and B_l are both contained in $\tilde{H}_{C_i C_m}$.

(b) For the remaining B_l that appear in combinations which have not yet been eliminated:

(i) For all $B_i \neq B_l$ in the shaded region, construct $\tilde{H}_{B_i B_l}$. If both C_m and A_k are contained in $\tilde{H}_{B_i B_l}$ for any B_l, eliminate the combination (A_k, B_l, C_m).

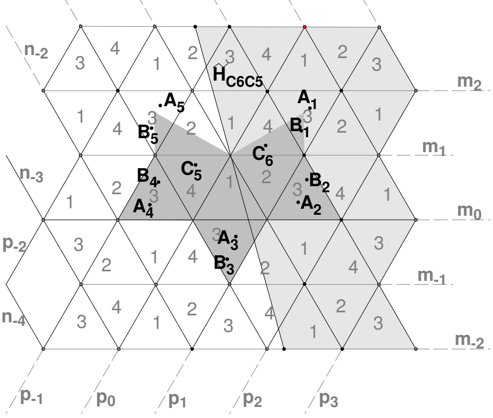
(ii) For the A_k that appear in a remaining combination with B_l and C_m: For all $A_i \neq A_k$ in the shaded region construct $\tilde{H}_{A_i A_k}$. If both C_m and B_l are contained in $\tilde{H}_{A_i A_k}$, eliminate the combination (A_k, B_l, C_m).

(4) Measure the lengths of the Steiner minimal trees formed from the remaining combinations using Algorithm 2.1. The Steiner minimal tree with shortest length realizes SMT(a, b, c). The algorithm is complete.

We will now demonstrate how to apply the algorithm for the configuration shown in Figure 9, which clearly corresponds to Case 3.

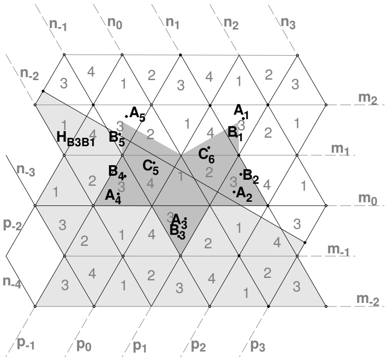
B_5, A_1 and A_5 are not contained within the shaded region, so none of (A_1, B_2, C_5), (A_1, B_3, C_6), (A_1, B_4, C_6), (A_5, B_4, C_5), (A_5, B_3, C_5), (A_3, B_5, C_5) and (A_4, B_5, C_5) need to be considered.

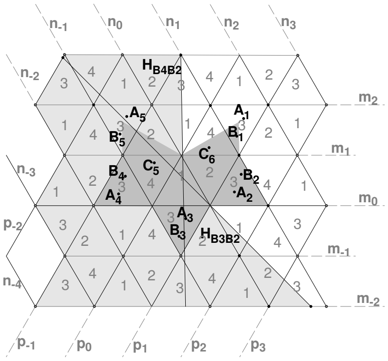
Construct $\tilde{H}_{C_6 C_5}$:



Since both A_2 and B_2 are contained in $\tilde{H}_{C_6C_5}$, the combination (A_2, B_2, C_5) can be eliminated.

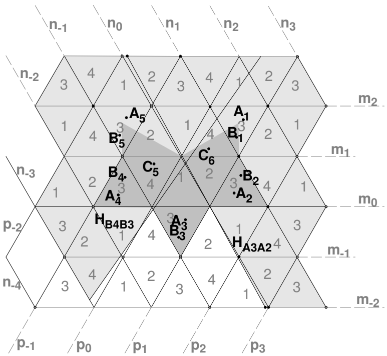
Construct $\tilde{H}_{B_iB_1}$ for all $i \neq 1$ (left diagram). Since C_5 and A_3 are contained in $\tilde{H}_{B_3B_1}$, the combination (A_3, B_1, C_5) can be eliminated. There are no remaining combinations which use both B_1 and C_5.

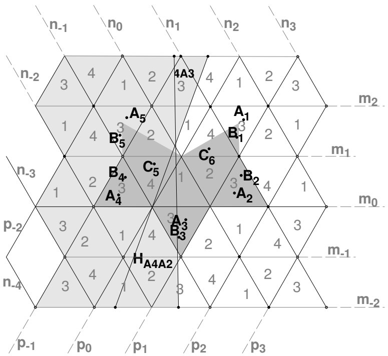




Construct $\tilde{H}_{B_iB_2}$ for all $i \neq 2$ (right diagram above). Since both C_5 and A_3 are contained in $\tilde{H}_{B_3B_2}$, the combination (A_3, B_2, C_5) can be eliminated. Since both C_5 and A_4 are contained in $\tilde{H}_{B_4B_2}$, the combination (A_4, B_2, C_5) can be eliminated. There are no remaining combinations which use both B_2 and C_5.

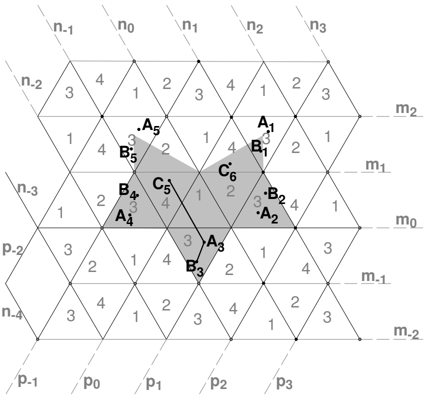
Construct $\tilde{H}_{B_iB_3}$ for all $i \neq 3$ (left diagram below). Since both C_5 and A_4 are contained in $\tilde{H}_{B_4B_3}$, the combination (A_4, B_3, C_5) can be eliminated. The only remaining combinations the list are (A_2, B_3, C_5) and (A_3, B_3, C_5). However, since both C_5 and B_3 are contained in $\tilde{H}_{A_3A_2}$, (A_2, B_3, C_5) can be eliminated.



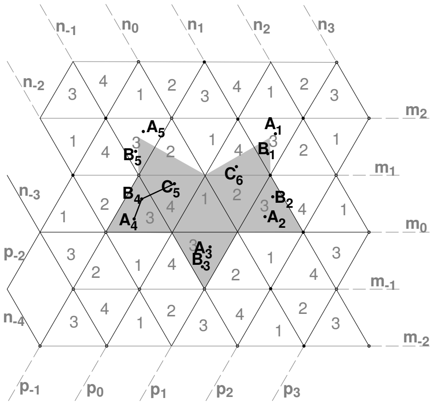


Construct $\tilde{H}_{B_iB_4}$ for all $i \neq 4$. Since C_5 is not contained in any $\tilde{H}_{B_iB_4}$ with $i \neq 4$, the remaining possibilities from the above list are (A_2, B_4, C_5), (A_3, B_4, C_5), and (A_4, B_4, C_5). Since both C_5 and B_4 are contained in $\tilde{H}_{A_4A_2}$, (A_2, B_4, C_5) can be eliminated (right diagram immediately above). Since both C_5 and B_4 are contained in $\tilde{H}_{A_4A_3}$, (A_3, B_4, C_5) can be eliminated.

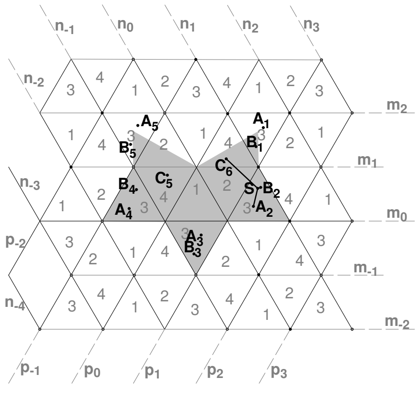
We have shown that the only remaining combinations in the list containing C_5 are (A_3, B_3, C_5) and (A_4, B_4, C_5). Using a similar procedure, we can show that the only remaining combination containing C_6 is (A_2, B_2, C_6). Assuming $\mathcal{T}$ has edge-length 1, we construct the Steiner trees associated with each of these combinations, with the following results:



$\mathcal{L}(\mathrm{SMT}(A_3, B_3, C_5)) = 1.04$



$\mathcal{L}(\mathrm{SMT}(A_4, B_4, C_5)) = 0.87$



$\mathcal{L}(\mathrm{SMT}(A_2, B_2, C_6)) = 1.43$

Hence, $\mathrm{SMT}(A_4, B_4, C_5)$ realizes $\mathrm{SMT}(a, b, c)$ with length 0.87, and the algorithm is complete with only three actual measurements.

References

[Brazil et al. 1998] M. Brazil, J. H. Rubinstein, D. A. Thomas, J. F. Weng, and N. C. Wormald, "Shortest networks on spheres", pp. 453–461 in *Network design: connectivity and facilities location* (Princeton, 1997), edited by P. M. Pardalos and D. Du, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. **40**, Amer. Math. Soc., Providence, RI, 1998. MR 98k:05046 Zbl 0915.05043

[Brune and Sipe 2009] T. Brune and L. Sipe, "Shortest path between two points on the regular tetrahedron", 2009, available at http://tinyurl.com/BruneSipe2009.

[Caffarelli et al. 2010] E. Caffarelli, D. M. Halverson, and R. J. Jensen, "The Steiner problem on surfaces of revolution", preprint, 2010.

[Gilbert and Pollak 1968] E. N. Gilbert and H. O. Pollak, "Steiner minimal trees", *SIAM J. Appl. Math.* **16** (1968), 1–29. MR 36 #6317 Zbl 0159.22001

[Halverson and March 2005] D. Halverson and D. March, "Steiner tree constructions in hyperbolic space", 2005, available at http://tinyurl.com/HalversonMarch2005.

[Halverson and Penrod 2007] D. Halverson and K. Penrod, "The three point Steiner problem on the flat torus", 2007, available at http://tinyurl.com/HalversonPenrod2007.

[Hwang and Weng 1986] F. K. Hwang and J. F. Weng, "Hexagonal coordinate systems and Steiner minimal trees", *Discrete Math.* **62**:1 (1986), 49–57. MR 87k:05064 Zbl 0601.05016

[Hwang et al. 1992] F. K. Hwang, D. S. Richards, and P. Winter, *The Steiner tree problem*, Annals of Discrete Mathematics **53**, North-Holland, Amsterdam, 1992. MR 94a:05051 Zbl 0774.05001

[Ivanov and Tuzhilin 1994] A. O. Ivanov and A. A. Tuzhilin, *Minimal networks: The Steiner problem and its generalizations*, CRC Press, Boca Raton, FL, 1994. MR 95h:05050 Zbl 0842.90116

[Kuhn 1974] H. W. Kuhn, " "Steiner's" problem revisited", pp. 52–70 in *Studies in optimization* (Washington, DC), edited by G. B. Dantzig and B. C. Eaves, Studies in Math. **10**, Math. Assoc. Amer., 1974. MR 57 #18835

[Lee et al. 2011] A. Lee, J. Lytle, D. Halverson, and D. Sampson, "The Steiner problem on narrow and wide cones", 2011, available at http://tinyurl.com/LeeLytleHalversonSampson2011.

[Litwhiler and Aly 1980] D. W. Litwhiler and A. A. Aly, "Steiner's problem and Fagnano's result on the sphere", *Math. Programming* **18**:3 (1980), 286–290. MR 81g:90082 Zbl 0433.90029

[May and Mitchell 2007] K. L. May and M. A. Mitchell, "The three point Steiner problem on the flat torus: the minimal lune case", 2007, available at http://tinyurl.com/MayMitchell2007.

[Melzak 1961] Z. A. Melzak, "On the problem of Steiner", *Canad. Math. Bull.* **4** (1961), 143–148. MR 23 #A2767 Zbl 0101.13201

[Weng 2001] J. F. Weng, "Steiner trees on curved surfaces", *Graphs Combin.* **17**:2 (2001), 353–363. MR 2003c:05058 Zbl 0982.05036

[Zacharias 1914–1921] M. Zacharias, "Elementargeometrie und elementare nicht-euklidische Geometrie in synthetische Behandlung", Article III AB.9 occupying part of Heft 5 (1914) and all of Heft 6 (1921) in *Encyklopädie der Mathematischen Wissenschaften*, vol. III (Geometrie), edited by W. F. Meyer and H. Mohrmann, Teubner, Leipzig.

Received: 2011-01-14 Revised: 2011-03-22 Accepted: 2011-03-24

kyrammoon@gmail.com
Mathematics Department, Brigham Young University,
275 TMCB, Provo, UT 84602, United States

s_gmshero@clarion.edu
Mathematics Department, Clarion University of Pennsylvania,
189 STC, 840 Wood Street, Clarion, PA 16214, United States

halverson@math.byu.edu
Mathematics Department, Brigham Young University,
263 TMCB, Provo, UT 84602, United States

INVOLVE 4:4(2011)

Constructions of potentially eventually positive sign patterns with reducible positive part

Marie Archer, Minerva Catral, Craig Erickson, Rana Haber,
Leslie Hogben, Xavier Martinez-Rivera and Antonio Ochoa

(Communicated by Chi-Kwong Li)

Potentially eventually positive (PEP) sign patterns were introduced by Berman et al. (*Electron. J. Linear Algebra* **19** (2010), 108–120), where it was noted that a matrix is PEP if its positive part is primitive, and an example was given of a 3×3 PEP sign pattern with reducible positive part. We extend these results by constructing $n \times n$ PEP sign patterns with reducible positive part, for every $n \geq 3$.

1. Introduction

A *sign pattern matrix* (or *sign pattern*) is a matrix having entries in $\{+, -, 0\}$. For a real matrix A, $\operatorname{sgn}(A)$ is the sign pattern having entries that correspond to the signs of the entries in A. If $\mathscr{A}$ is an $n \times n$ sign pattern, the *qualitative class* of $\mathscr{A}$, denoted $Q(\mathscr{A})$, is the set of all $A \in \mathbb{R}^{n\times n}$ such that $\operatorname{sgn}(A) = \mathscr{A}$, where $\operatorname{sgn}(A) = [\operatorname{sgn}(a_{ij})]$; such a matrix A is called a *realization* of $\mathscr{A}$. Qualitative matrix problems were introduced by Samuelson [1947] in the mathematical modeling of problems from economics. Sign pattern matrices have useful applications in economics, population biology, chemistry and sociology. If P is a property of a real matrix, then a sign pattern $\mathscr{A}$ is *potentially* P (or *allows* P) if there is some $A \in Q(\mathscr{A})$ that has property P.

The *spectrum* of a square matrix A, denoted $\sigma(A)$, is the multiset of the eigenvalues of A, and the *spectral radius* of A is defined as $\rho(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}$. Matrix A has the *strong Perron–Frobenius property* if $\rho(A) > 0$ is a simple strictly dominant eigenvalue of A that has a positive eigenvector. A matrix $A \in \mathbb{R}^{n\times n}$ is *eventually positive* if there exists a $k_0 \in \mathbb{Z}^+$ such that for all $k \geq k_0$, $A^k > 0$, where the inequality is entrywise. Handelman developed the following test for eventual positivity in [Handelman 1981]: a matrix A is eventually positive if and only if both A and A^T satisfy the strong Perron–Frobenius property. If there exists a k such

MSC2010: 15B35, 15B48, 05C50, 15A18.
Keywords: potentially eventually positive, PEP, sign pattern, matrix, digraph.

that $A^k > 0$ and $A^{k+1} > 0$, then A is eventually positive [Johnson and Tarazaga 2004]. A sign pattern $\mathscr{A}$ is *potentially eventually positive* (PEP) if there exists an eventually positive realization $A \in Q(\mathscr{A})$.

For a sign pattern $\mathscr{A} = \left[\alpha_{ij}\right]$, define the *positive part* of $\mathscr{A}$ to be $\mathscr{A}^+ = \left[\alpha_{ij}^+\right]$ and the *negative part* of $\mathscr{A}$ to be $\mathscr{A}^- = \left[\alpha_{ij}^-\right]$, where

$$\alpha_{ij}^+ = \begin{cases} + & \text{if } \alpha_{ij} = +, \\ 0 & \text{if } \alpha_{ij} = 0 \text{ or } \alpha_{ij} = -, \end{cases} \qquad \alpha_{ij}^- = \begin{cases} - & \text{if } \alpha_{ij} = -, \\ 0 & \text{if } \alpha_{ij} = 0 \text{ or } \alpha_{ij} = +. \end{cases}$$

Clearly $\mathscr{A} = \mathscr{A}^+ + \mathscr{A}^-$. For a matrix $A \in \mathbb{R}^{n\times n}$, the positive part A^+ of A and negative part A^- of A are defined analogously, and $A = A^+ + A^-$.

A *digraph* $\Gamma = (V, E)$ consists of a finite, nonempty set V of vertices, together with a set $E \subseteq V \times V$ of arcs. Note that a digraph allows loops (arcs of the form (v, v)) and may have both arcs (v, w) and (w, v) but not multiple copies of the same arc. Let $A = \left[a_{ij}\right] \in \mathbb{R}^{n\times n}$. The *digraph of* A, denoted $\Gamma(A)$, has vertex set $\{1, \dots, n\}$ and arc set $\left\{(i, j) : a_{ij} \neq 0\right\}$. If $\mathscr{A}$ is a sign pattern, then $\Gamma(\mathscr{A}) = \Gamma(A)$ where $A \in Q(\mathscr{A})$. A digraph Γ is *strongly connected* if for any two distinct vertices v and w of Γ, there is a path in Γ from v to w.

A square matrix A is *reducible* if there exists a permutation matrix P such that

$$PAP^T = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}$$

where A_{11} and A_{22} are nonempty square matrices and 0 is a (possibly rectangular) block consisting entirely of zero entries, or A is the 1×1 zero matrix. If A is not reducible, then A is called *irreducible*. It is well known that for $n \geq 2$, A is irreducible if and only if $\Gamma(A)$ is strongly connected. For a strongly connected digraph Γ, the *index of imprimitivity* is the greatest common divisor of the lengths of the cycles in Γ. A strongly connected digraph is *primitive* if its index of imprimitivity is one; otherwise it is *imprimitive*. The *index of imprimitivity* of a nonnegative sign pattern $\mathscr{A}$ is the index of imprimitivity of $\Gamma(\mathscr{A})$ and $\mathscr{A} \geq 0$ is *primitive* if $\Gamma(\mathscr{A})$ is primitive, or equivalently, if the index of imprimitivity of $\mathscr{A}$ is one.

The study of PEP sign patterns was introduced in [Berman et al. 2010], where it was shown that if $\mathscr{A}^+$ is primitive, then $\mathscr{A}$ is PEP, and where the first example of a PEP sign pattern with reducible positive part was given: the 3×3 pattern

$$\mathscr{B} = \begin{bmatrix} + & - & 0 \\ + & 0 & - \\ - & + & + \end{bmatrix}.$$

In Section 2 we extend the results of [Berman et al. 2010] by generalizing the 3×3 pattern $\mathscr{B}$ given there to a family of PEP sign patterns having reducible positive part for every order $n \geq 3$.

In Section 3 we examine the effect of the Kronecker product on PEP sign patterns and obtain another method of constructing PEP sign patterns with reducible positive part.

2. A family of sign patterns generalizing $\mathcal{B}$

The sign pattern $\mathcal{B}$ from [Berman et al. 2010] was the first PEP sign pattern with a reducible positive part. This sign pattern may be generalized by defining the $n \times n$ sign pattern

$$\mathcal{B}_n = \begin{bmatrix} + & - & \cdots & - & 0 \\ + & 0 & \cdots & 0 & - \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ + & 0 & \cdots & 0 & - \\ - & + & \cdots & + & + \end{bmatrix}.$$

The following result, which is a special case of the *Schur–Cohn criterion* (see, e.g., [Marden 1949]), will be used in the proof that $\mathcal{B}_n$ is PEP.

Lemma 2.1. *If the polynomial $f(x) = x^2 - \beta x + \alpha$ satisfies $|\beta| < 1 + \alpha < 2$, then all zeros of $f(x)$ lie strictly inside the unit circle.*

It is well known that if the characteristic polynomial of A is $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ then $a_{n-k} = (-1)^k E_k(A)$, where $E_k(A)$ is the sum of the $k \times k$ principal minors of A (see, e.g., [Horn and Johnson 1985]).

Theorem 2.2. *For $n \geq 3$ the $n \times n$ sign pattern $\mathcal{B}_n$ is PEP.*

Proof. For $t > 0$, let $B_n(t)$ be the $n \times n$ matrix

$$B_n(t) = \begin{bmatrix} 1+(n-2)t & -t & \cdots & -t & 0 \\ 1+t & 0 & \cdots & 0 & -t \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1+t & 0 & \cdots & 0 & -t \\ -(n-2)t - \frac{1}{2}t^2 & t & \cdots & t & 1+\frac{1}{2}t^2 \end{bmatrix}.$$

Then $B_n(t) \in Q(\mathcal{B}_n)$, and 1 is an eigenvalue of $B_n(t)$ with positive right eigenvector $\mathbb{1}$ (the all ones vector) and positive left eigenvector

$$\boldsymbol{w} = \begin{bmatrix} \frac{2n-5}{t} & 1 & \cdots & 1 & \frac{2n-4}{t} \end{bmatrix}^T.$$

We show that for some choice of $t > 0$, 1 is a simple strictly dominant eigenvalue of $B_n(t)$ and hence $B_n(t)$ is eventually positive. Since $1 \in \sigma(B_n(t))$ and rank $B_n(t) \leq 3$, the characteristic polynomial $p_{B_n(t)}(x)$ of $B_n(t)$ is of the form

$$p_{B_n(t)}(x) = x^{n-3}(x-1)(x^2 - \beta x + \alpha) = x^n - (1+\beta)x^{n-1} + (\alpha+\beta)x^{n-2} - \alpha x^{n-3}.$$

Computing α and β using the sums of principal minors to evaluate the characteristic polynomial gives $\beta = \frac{1}{2}t^2 + (n-2)t + 1$ and $\alpha = (n-2)t\left(1 + 2t + \frac{1}{2}t^2\right)$. For $n > 3$, setting $t = 1/(2(n-2))$ gives $|\beta| < 1 + \alpha < 2$, which, using Lemma 2.1, guarantees that the two nonzero eigenvalues of B_n other than 1 have modulus strictly less than 1 (recall that a 3×3 eventually positive matrix $B_3 \in Q(\mathcal{B}_3)$ was given in [Berman et al. 2010] so we have not been concerned with this case in choosing t). □

We illustrate this theorem with an example.

Example 2.3. Let $n = 5$. Following the proof of Theorem 2.2, we choose $t = \frac{1}{6}$ and define

$$B_5 = B_5\left(\frac{1}{6}\right) = \frac{1}{6}\begin{bmatrix} 9 & -1 & -1 & -1 & 0 \\ 7 & 0 & 0 & 0 & -1 \\ 7 & 0 & 0 & 0 & -1 \\ 7 & 0 & 0 & 0 & -1 \\ -\frac{37}{12} & 1 & 1 & 1 & \frac{73}{12} \end{bmatrix}.$$

Moreover, we have

$$\begin{aligned} \sigma(B_5) &= \left\{1, \tfrac{1}{144}\left(109 + i\sqrt{2087}\right), \tfrac{1}{144}\left(109 - i\sqrt{2087}\right), 0, 0\right\} \\ &\approx \{1,\ 0.7569 + 0.3172i,\ 0.7569 - 0.3172i,\ 0,\ 0\}, \end{aligned}$$

and $\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}^T$ and $\begin{bmatrix} \frac{5}{6} & \frac{1}{36} & \frac{1}{36} & \frac{1}{36} & 1 \end{bmatrix}^T$ are right and left eigenvectors, respectively, corresponding to $\rho(B_5) = 1$. Therefore B_5 and B_5^T have the strong Perron–Frobenius property, so B_5 is eventually positive by Handelman's criterion.

In [Berman et al. 2010] it was shown that if the sign pattern $\mathcal{A}$ is PEP, then any sign pattern achieved by changing one or more zero entries of $\mathcal{A}$ to be nonzero is also PEP. Applying this to $\mathcal{B}_n$ yields a variety of additional PEP sign patterns having reducible positive part.

3. Kronecker products

The Kronecker product (sometimes called the tensor product) is a useful tool for generating larger eventually positive matrices and thus PEP sign patterns. The *Kronecker product* of $A = [a_{ij}]$ and $B = [b_{ij}]$ is defined as

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{n1}B & \cdots & a_{nn}B \end{bmatrix}.$$

It is clear that if $A > 0$ and $B > 0$, then $A \otimes B > 0$. The following facts can be found in many linear algebra books; see [Reams 2006], for example. For $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$, $(A \otimes B)^k = A^k \otimes B^k$. For A, C, B, D of appropriate dimensions, we have $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$. There exists a permutation matrix P such that $B \otimes A = P(A \otimes B)P^T$.

Proposition 3.1. *If A and B are eventually positive matrices, then $A \otimes B$ is eventually positive.*

Proof. Assume that A and B are eventually positive matrices. Since A and B are eventually positive, there exists some $s_0, t_0 \in \mathbb{Z}$, with $s_0, t_0 > 0$, such that for all $s \geq s_0$ and $t \geq t_0$, $A^s > 0$ and $B^t > 0$. Set $k_0 = \max\{s_0, t_0\}$. Then for all $k \geq k_0$, $(A \otimes B)^k = A^k \otimes B^k > 0$. □

Corollary 3.2. *If $\mathcal{A}$ and $\mathcal{B}$ are PEP sign patterns, then $\mathcal{A} \otimes \mathcal{B}$ is PEP.*

If either A or B is a reducible matrix, then $A \otimes B$ is reducible since, without loss of generality, if

$$PAP^T = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}$$

then

$$(P \otimes I)(A \otimes B)(P \otimes I)^T = \begin{bmatrix} A_{11} \otimes B & 0 \\ A_{21} \otimes B & A_{22} \otimes B \end{bmatrix}.$$

Thus Corollary 3.2 provides another way to construct PEP sign patterns having reducible positive part.

Example 3.3. Let

$$B = \frac{1}{100} \begin{bmatrix} 130 & -30 & 0 \\ 130 & 0 & -30 \\ -31 & 30 & 101 \end{bmatrix}.$$

In [Berman et al. 2010] it was shown that B is eventually positive, and in fact $B^k > 0$ for $k \geq 10$.

Let $A = \left[\begin{smallmatrix} 2 & 3 \\ 1 & 0 \end{smallmatrix}\right]$. Then $A^k > 0$ for $k \geq 2$, hence A is eventually positive.

Then

$$B \otimes A = \frac{1}{100} \begin{bmatrix} 260 & 390 & -60 & -90 & 0 & 0 \\ 130 & 0 & -30 & 0 & 0 & 0 \\ 260 & 390 & 0 & 0 & -60 & -90 \\ 130 & 0 & 0 & 0 & -30 & 0 \\ -62 & -93 & 60 & 90 & 202 & 303 \\ -31 & 0 & 30 & 0 & 101 & 0 \end{bmatrix}.$$

Moreover $(B \otimes A)^{10} > 0$ and $(B \otimes A)^{11} > 0$, so $B \otimes A$ is eventually positive and $\operatorname{sgn}(B \otimes A)$ is a PEP sign pattern with reducible positive part.

Any 0 in $\operatorname{sgn}(B \otimes A)$ from Example 3.3 may be changed to $-$ to get yet another PEP sign pattern with reducible positive part.

References

[Berman et al. 2010] A. Berman, M. Catral, L. M. DeAlba, A. Elhashash, F. J. Hall, L. Hogben, I.-J. Kim, D. D. Olesky, P. Tarazaga, M. J. Tsatsomeros, and P. van den Driessche, "Sign patterns that allow eventual positivity", *Electron. J. Linear Algebra* **19** (2010), 108–120. MR 2011c:15089 Zbl 1190.15031

[Handelman 1981] D. Handelman, "Positive matrices and dimension groups affiliated to C^*-algebras and topological Markov chains", *J. Operator Theory* **6**:1 (1981), 55–74. MR 84i:46058 Zbl 0495.06011

[Horn and Johnson 1985] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, 1985. MR 87e:15001 Zbl 0576.15001

[Johnson and Tarazaga 2004] C. R. Johnson and P. Tarazaga, "On matrices with Perron–Frobenius properties and some negative entries", *Positivity* **8**:4 (2004), 327–338. MR 2005k:15020 Zbl 1078.15018

[Marden 1949] M. Marden, *The geometry of the zeros of a polynomial in a complex variable*, Mathematical Surveys **3**, American Mathematical Society, Providence, RI, 1949. 2nd ed. titled *Geometry of polynomials* in 1966. MR 37 #1562 Zbl 0162.37101

[Reams 2006] R. Reams, "Partitioned matrices", Chapter 10, in *Handbook of linear algebra*, edited by L. Hogben, Chapman & Hall/CRC, Boca Raton, FL, 2006. MR 2007j:15001 Zbl 1122.15001

[Samuelson 1947] P. A. Samuelson, *Foundations of economic analysis*, Harvard University Press, Cambridge, MA, 1947. MR 10,555b Zbl 0031.17401

Received: 2011-03-03 Accepted: 2011-06-10

mharcher@iastate.edu

*Department of Mathematics,
Iowa State University of Science and Technology,
396 Carver Hall, Ames, IA 50011-2064, United States*

*Department of Mathematics, Columbia College,
Columbia, SC 29203, United States*

catralm@xavier.edu

*Department of Mathematics and Computer Science,
Xavier University, Cincinnati, OH 45207, United States*

craig@iastate.edu

*Department of Mathematics,
Iowa State University of Science and Technology,
396 Carver Hall, Ames, IA 50011-2064, United States*

rhaber2010@my.fit.edu

*Mathematics Department, Florida Institute of Technology,
Melbourne, FL 32901, United States*

lhogben@iastate.edu

*Department of Mathematics,
Iowa State University of Science and Technology,
396 Carver Hall, Ames, IA 50011-2064, United States*

*American Institute of Mathematics, 360 Portage Avenue,
Palo Alto, CA 94306, United States*

xavier.martinez@upr.edu

*Department of Mathematical Sciences, University of Puerto
Rico, Mayagüez, P.R. 00681, United States*

aochoa@csupomona.edu

*California State Polytechnic University, Pomona,
Pomona, CA 91768, United States*

INVOLVE 4:4(2011)

Congruence properties of S-partition functions

Andrew Gruet, Linzhi Wang, Katherine Yu and Jiangang Zeng

(Communicated by Ken Ono)

We study the function $p(S; n)$ that counts the number of partitions of n with elements in S, where S is a set of integers. Generalizing previous work of Kronholm, we find that given a positive integer m, the coefficients of the generating function of $p(S; n)$ are periodic modulo m, and we use this periodicity to obtain families of S-partition congruences. In particular, we obtain families of congruences between partition functions $p(S_1; n)$ and $p(S_2; n)$.

1. Introduction and statement of results

The *partition function* $p(n)$ is the number of nonincreasing sequences of positive integers that sum to n. Ramanujan proved the following congruences for $p(n)$:

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

Let S be a finite set of positive integers. An *S-partition* of an integer n is any nonincreasing sequence of integers in S that sums to n. The S-partition function $p(S; n)$ counts the number of S-partitions of n. The generating function for $p(S; n)$ is

$$G(S; q) := \sum_{n=0}^{\infty} p(S; n) q^n = \frac{1}{\prod_{s \in S}(1-q^s)} \in \mathbb{Z}[\![q]\!]. \tag{1-1}$$

Kronholm [2005; 2007] found elegant "Ramanujan-type" congruences for the partition function

$$p(n, m) = p(\{1, \dots, m\}; n-m).$$

In this paper we reinterpret his idea of periodicity and we generalize it in the context of sets of positive integers. We first show that the coefficients of the generating function above are periodic modulo m.

MSC2010: 11P83.

Keywords: Brandt Kronholm, Ramanujan-type congruences, S-partition functions.

Theorem 1.1. *For a finite set of positive integers S and a positive integer m, there exists a positive integer $\gamma_m(S)$ such that for every integer n and all nonnegative k, we have*

$$p(S;n)\equiv p(S;n+k\gamma_m(S)) \pmod{m}. \tag{1-2}$$

Example. This theorem immediately implies many Ramanujan-type congruences. For example, if $S=\{1,2,3,5\}$, one easily verifies that $\gamma_7(S)=210$. Therefore, the fact that $p(S;20)=91\equiv 0 \pmod 7$ gives the Ramanujan congruence

$$p(S;210n+20)\equiv 0 \pmod{7}.$$

Example. This theorem is analogous to Theorem 2 of [Kronholm 2007]. As Kronholm states, let d be a multiple of $\operatorname{lcm}\{1,\dots,t\}$ and for the odd prime m, let m^α be a primary factor of d. Kronholm shows that if we let $\gamma_m(\{1,\dots,t\}):=d$, the congruences (1-2) hold. In particular, he proves that if

$$\sum_{\delta\ge 0} m^\delta\left(\left\lfloor\frac{t}{m^\delta}\right\rfloor-\left\lfloor\frac{t}{m^{\delta+1}}\right\rfloor\right)\le m^\alpha, \tag{1-3}$$

then for $n\ge d-\sum_{j=2}^{t-1} j$, we have

$$p(\{1,\dots,t\};n-t)\equiv p(\{1,\dots,t\};n-t-d) \pmod{m}. \tag{1-4}$$

Example. Theorem 1.1 extends Theorem 2 of [Kronholm 2007] in that m does not have to be an odd prime. Let $S:=\{2,3,11\}$ and let $m:=12$. Given this choice of S, it is clear that $p(S;1)=0$. We find that $\gamma_{12}(S)=792$. By our theorem, for all positive k, $p(S;1+792k)\equiv 0 \pmod{12}$.

For convenience, for sets S we let

$$\Phi_S(q):=\prod_{s\in S}(1-q^s). \tag{1-5}$$

Corollary 1.2. *Let S_1 and S_2 be finite sets of positive integers and let m be a positive integer. If $\Phi_{S_1}(q)$ divides $\Phi_{S_2}(q)$ in $(\mathbb{Z}/m\mathbb{Z})[q]$, then for any nonnegative integer d, let*

$$X(q):=q^{d\gamma_m(S_1)}\frac{\Phi_{S_2}(q)}{\Phi_{S_1}(q)},$$

where

$$q^{d\gamma_m(S_1)}\frac{\Phi_{S_2}(q)}{\Phi_{S_1}(q)}:\equiv\sum_{i=0}^{c}a_iq^i \pmod{m}.$$

For $n\ge c$ and for any nonnegative k_1 and k_2, we have

$$p(S_1;n+k_1\gamma_m(S_1))\equiv\sum_{i=0}^{c}a_i\,p(S_2;n-i+k_2\gamma_m(S_2)) \pmod{m}. \tag{1-6}$$

Note that Corollary 1.2 applies for all m when $S_1 \subseteq S_2$.

Example. Let $S_1 := \{1, 17\}$ and $S_2 := \{17, 289\}$ and let $m := 17$. Let $X(q) := \Phi_{S_2}(q)/\Phi_{S_1}(q)$. Then clearly $X(q) = \sum_{i=0}^{288} q^i$. For all nonnegative k_1 and k_2 and for all $n \geq 288$,

$$p(S_1; n + 289k_1) \equiv \sum_{i=0}^{288} p(S_2; n - i + 4913k_2) \pmod{17}.$$

2. Proof of Theorem 1.1

For convenience, we let $S := \{s_1, s_2, \dots, s_t\}$ and let $d_S = \sum_{i=1}^{t} s_i$. Let

$$\Phi_S(q) = \prod_{s \in S} (1 - q^s) := \sum_{n=0}^{d_s} b(S; n) q^n.$$

From the identity

$$1 = \Phi_S(q) G(S; q) = \left(\sum_{n=0}^{d_s} b(S; n) q^n \right) \left(\sum_{n=0}^{\infty} p(S; n) q^n \right),$$

we have

$$1 = \sum_{i \geq 0} b(S; 0) p(S; i) q^i + \sum_{i \geq 0} b(S; 1) p(S; i) q^{i+1} + \cdots + \sum_{i \geq 0} b(S; d_S) p(S; i) q^{i+d_S}.$$

Looking at coefficients of q^N for $N \geq 1$, we observe that

$$\sum_{n=0}^{d_S} b(S; n) p(S; N - n) = 0. \tag{2-1}$$

This defines a linear recurrence relation. Noting that $b(S; 0) = 1$, we have

$$p(S; N) = - \sum_{n=1}^{d_S} b(S; n) p(S; N - n).$$

We consider consecutive d_S-tuples of consecutive partition values. Arranging these tuples in order, we have

$$\begin{gathered} \big(p(S; 0), p(S; 1), \dots, p(S; d_S - 1)\big), \\ \big(p(S; d_S), p(S; d_S + 1), \dots, p(S; 2d_S - 1)\big), \end{gathered}$$

and so on. By reducing modulo m, we will find a first pair of d_S-tuples that agrees modulo m. Indeed, the maximal possible number of different tuples is m^{d_S}. Suppose that the first tuple of this pair starts at $p(S; n_0)$, and the second tuple starts at $p(S; n_1)$. Since by the linear recurrence relation, each tuple determines the next tuple, we have inductively that for all nonnegative i,

$$p(S; n_0 + i) \equiv p(S; n_1 + i) \pmod{m}.$$

We will first show that the residue classes of each tuple after the first determines the preceding tuple's residue classes. For any $a_0 = vd_S$, with $v \geq 1$, we consider the tuple

$$\big(p(S; a_0), \ldots, p(S; a_0 + d_S - 1)\big).$$

By (2-1), and noting that $b(S; d_S) = (-1)^t$, we have

$$(-1)^{t+1} p(S; N - d_S) = \sum_{n=0}^{d_S - 1} b(S; n) p(S; N - n).$$

It follows immediately that

$$\begin{aligned}
(-1)^{t+1} p(S; a_0 - 1) &\equiv \sum_{i=0}^{d_S - 1} b(S; i) p(S; a_0 + d_S - 1 - i) \pmod{m},\\
(-1)^{t+1} p(S; a_0 - 2) &\equiv \left(\sum_{i=0}^{d_S - 2} b(S; i) p(S; a_0 + d_S - 2 - i) \right)\\
&\qquad\qquad + b(S; d_S - 1) p(S; a_0 - 1) \pmod{m},\\
&\;\;\vdots\\
(-1)^{t+1} p(S; a_0 - d_S) &\equiv b(S; 0) p(S; a_0) + \sum_{i=1}^{d_S - 1} b(S; i) p(S; a_0 - i) \pmod{m}.
\end{aligned}$$

Therefore, the residue classes of $\big(p(S; a_0), \ldots, p(S; a_0 + d_S - 1)\big)$ reduced modulo m uniquely determine the residue classes of $\big(p(S; a_0 - d_S), \ldots, p(S; a_0 - 1)\big)$ reduced modulo m.

To complete the proof, we must show that $n_0 = 0$. By hypothesis,

$$\big(p(S; n_0), \ldots, p(S; d_S - 1)\big) \equiv \big(p(S; n_1), \ldots, p(S; n_1 + d_S - 1)\big) \pmod{m}.$$

Suppose $n_0 = vd_S$ where $v \geq 1$, (i.e., $n_0 \neq 0$). Then, by the argument above, we have

$$\big(p(S; n_0 - d_S), \ldots, p(S; n_0 - 1)\big) \equiv \big(p(S; n_1 - d_S), \ldots, p(S; n_1 - 1)\big) \pmod{m}.$$

This result contradicts our hypothesis that the first-repeated tuple started for the first time at $p(S; n_0)$. Therefore, we can conclude that $n_0 = 0$, and so we let $\gamma_m(S) := n_1$. In particular, for any nonnegative k, we have (1-2).

3. Proof of Corollary 1.2

By Theorem 1.1, for any nonnegative k_1, we have

$$p(S_1; n + k_1\gamma_m(S_1)) \equiv p(S_1; n) \pmod{m}.$$

Clearly, for any nonnegative k_2, we have

$$\sum_{i=0}^{c} a_i\, p(S_2; n - i + k_2\gamma_m(S_2)) \equiv \sum_{i=0}^{c} a_i\, p(S_2; n - i) \pmod{m}.$$

Thus, subtracting two congruences, we have (1-6):

$$p(S_1; n + k_1\gamma_m(S_1)) - \sum_{i=0}^{c} a_i\, p(S_2; n - i + k_2\gamma_m(S_2)) \equiv p(S_1; n) - \sum_{i=0}^{c} a_i\, p(S_2; n - i) \pmod{m}. \tag{3-1}$$

Since $\dfrac{\Phi_{S_2}(q)}{\Phi_{S_1}(q)} G(S_2; q) = G(S_1; q)$, we know

$$G(S_1; q) - X(q)G(S_2; q) \equiv G(S_1; q) - q^{d\gamma_m(S_1)} G(S_1; q) \pmod{m}. \tag{3-2}$$

By comparing coefficients in (3-2), we have, for $n \geq d\gamma_m(S_1)$,

$$p(S_1; n) - \sum_{i=0}^{c} a_i\, p(S_2; n - i) \equiv 0 \pmod{m}. \tag{3-3}$$

Thus by (3-1), if $n \geq c$, then for any nonnegative k_1 and k_2, since $c \geq d\gamma_m(S_1)$, we have

$$p(S_1; n + k_1\gamma_m(S_1)) \equiv \sum_{i=0}^{c} a_i\, p(S_2; n - i + k_2\gamma_m(S_2)) \pmod{m}.$$

4. Acknowledgements

The authors thank Ken Ono, Marie Jameson and Zach Kent for running the research course at Emory University in which this paper was written.

References

[Kronholm 2005] B. Kronholm, "On congruence properties of $p(n, m)$", *Proc. Amer. Math. Soc.* **133**:10 (2005), 2891–2895. MR 2006e:11161 Zbl 1065.05014

[Kronholm 2007] B. Kronholm, "On congruence properties of consecutive values of $p(n, m)$", *Integers* **7**:1 (2007), #A16. MR 2008c:11138 Zbl 05139414

Received: 2011-04-28 Accepted: 2011-06-17

agruet@emory.edu
PO Box 2054, Acton, MA 01720, United States

lwang75@emory.edu
No. 6 Building 14, Fuhehuayuan, Shuliang County,
Chihingdu, Sichuan, China

yu.katherin@gmail.com
51 Clearfield Dr., San Francisco, CA 94132, United States

jzeng6@emory.edu
1503 North Decatur Rd, No. 4, Atlanta 30327, United States

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@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

2011 vol. 4 no. 4

1944-4176(2011)4:4;1-B