Research Experience for Undergraduates
Research Reports

Indiana University, Bloomington

Summer 2013

Contents

Preface

During the summer of 2013 thirteen students participated in the Research Experiences for Undergraduates program in Mathematics at Indiana University. The program ran for eight weeks, from June 3 through July 26, 2013. Six faculty served as research advisers:

- Richard Bradley worked with Ariana Coppess (IU) and Emily Walther (Westminster)
- Marlies Gerber worked with Tom Dauer (IU)
- Evie Malaia (U. Texas at Arlington), assisted by IU grad student Jonathan Poelhuis, worked with Erik Bates (Michigan State), Katherine Coppess (U. Michigan), and Ben Seitzman (IU) on our first joint mathematics-neuroscience project
- William Orrick worked with Yancy Liao (Penn State) and Wade Bloomquist (U. Iowa) on two separate projects
- Kevin Pilgrim worked with Nicholas Neuctherlein (U. Michigan) and Samantha Pinella (Edinburgh) on one project, and with Nicholas Miller (U. Missouri-Columbia) and Max Zhou (IUB) on another
- Sidney Shaw (biology) worked with Allison Brumfield (St. Olaf College).

The program opened with an introductory pizza party. On the following morning, students began meeting with their faculty mentors; these meetings continued regularly throughout the first few weeks. During week one, there were short presentations by faculty mentors briefly introducing the problem to be investigated. Several other IU faculty gave talks on their favorite topics during the first half of the program. Students also received an orientation to the mathematics library. Week two featured individualized workshops in LaTex, run by graduate student Anne Carter. In week three, students attended a workshop regarding ethics in the profession, and students gave short, informal presentations to each other on the status of work on the project. They also attended a tour of the functional MRI brain imaging facility and EEG lab in the Department of Psychological and Brain Sciences, led by REU participant, Goldwater scholar, math and neuroscience major Ben Seitzman. Week four featured a tour of the puzzle collection at the Lilly Library, a campuswide reception for REU programs at the IMU faculty club, and a self-guided tour of the Morton Bradley sculptures in the nearby Mauer School of Law Library and IU Art Museum. In week five, they received a tour of the Center for the Exploration of Energy and Matter (cyclotron) facility led by Prof. Baxter, and attended a lecture on materials science there. During week six, they attended a pool party at local Bryan Park, hosted by Professor Elizabeth Housworth. During week eight, we hosted the Indiana Mathematics Undergraduate Research conference, which featured 21 lectures by 40 students from Goshen College, IU,

IUPUI, U. Notre Dame, Purdue University, and Valparaiso University. This concluded with a plenary lecture by Professor Rodrigo Perez of IUPUI on the theory of Siegel disks. The program concluded with a dinner at local eatery Max's Place and the submission of final reports, contained in this volume.

It took the help and support of many different groups and individuals to make the program a success.

We thank the National Science Foundation for major financial support through the REU program through NSF grant DMS-1156515. Arianna Cappon was partially supported by the Women in Science program, led by Indiana University's Division of Student Affairs. Additional financial support for mentors was provided by the Department of Mathematics and fpr graduate students by Elizabeth Housworth's NSF grant DMS-1206405. We thank the staff of the Department of Mathematics for support, especially Mandie McCarty for coordinating the complex logistical arrangments (housing, paychecks, information packets, meal plans, frequent shopping for snacks). We thank Indiana graduate student Anne Carter for serving as LaTeXconsultant and for compiling this volume.

Thanks to mathematics faculty Richard Bradley, Marlies Gerber, William Orrick, and Kevin Pilgrim for serving as mentors and giving lectures, to biology faculty member Sidney Shaw for serving as mentor and giving lectures, and to U Texas at Arlington faculty member Evie Malaia for serving as mentor and giving a lecture. We also thank IU mathematics faculty members Jiri Dadok, Elizabeth Housworth, Chris Judge, Bruce Solomon, Matthias Strauch, and Dylan Thurston for giving lectures to our group. Thanks to David Baxter of the Center for Exploration of Energy and Matter (nee IU cyclotron facility) for his personal tour of the cyclotron facility and lecture on the physics of materials. Thanks to Rebecca Bauman for her tour of the Slocum puzzle collection, and for gathering together for display and examination copies of Euler, Cauchy, and other authors in the original.

From left to right, rear: Max Zhou, Nicholas Miller, Katherine Coppess, Wade Bloomquist, Tom Dauer; front: Benjamin Seitzman, Arianna Cappon, Allison Brumfield, Emily Walther, Samantha Pinella, Erik Bates, Nicholas Neuchterline, Yancy Liao.

KMP
August 26, 2013

EEG Time Series Analysis and Functional Connectivity Network Measures of TD and ASD Youths*

Erik Bates† *Katherine Coppess*‡ *Benjamin Seitzman*§

Abstract

Graph theory allows for investigation of the arrangement and dynamics of connections between objects in a complex system. A prominent application of graph theory is featured in the study of neural networks of the human brain. Time series representations of brain activity are acquired from neuroimaging methodologies, such as electroencephalography (EEG). EEG records electric potential changes in global brain activity across the scalp as a function of time. Previously recorded 32-channel EEG data of typically developing (TD) youths and youths with Autism Spectrum Disorder (ASD) during both wakeful rest and a visual task were analyzed. A cross-correlation analysis of the EEG time series was used to produce weighted, undirected graphs corresponding to functional brain networks. The stability of these networks was assessed by novel use of the ℓ-1 norm for matrix entries, here called the edit distance. Upon examination of stable networks identified, there was a significantly larger number of stable networks observed in the resting condition compared to the task condition. Furthermore, stable networks were found to endure a significantly longer time during the resting condition in children with ASD than in TD children.

1 Introduction

Recently, neuroscientists have applied network science to the study of the brain using graph theory. Graph theory, at its most basic level, investigates the arrangement and the nature of connections between objects. When graph theory is applied to the brain, the objects are neurons (or groups of neurons), and the connections are the anatomical or functional links between them. Regarding the brain as a network has revolutionized researchers' attempts to understand both normal and abnormal brain function [2, 6, 13, 16, 18]. One dysfunction widely studied is Autism Spectrum Disorder (ASD), which is commonly characterized by atypical communication abilities, social understanding, and executive

*This research is supported in part by NSF grant DMS-1156515.

†Michigan State University, bateser2@msu.edu

‡University of Michigan, kcoppess@umich.edu

§Indiana University, beaseitz@indiana.edu

processing. Only recently has research of the functional differences in brain network activity between typically developing (TD) children and children with ASD progressed [1, 5, 7, 12, 19]. The functional dynamics of these networks may be captured via several neuroimaging modalities, including electroencephalography (EEG). EEG is a non-invasive method of recording the brain's electrical activity from the surface of the human scalp [11]. The high temporal resolution of EEG recordings allows for the examination of brain dynamics on the millisecond timescale. Brain activity is recorded as potential changes in time, thereby generating a number of time series that each correspond to one measurement site. The collected time series may be used to understand functional networks of the brain and, in particular, how those networks evolve over time. One topic of interest in the neuroscience community is if and when such networks remain quasi-stable [3, 8, 9].

The question of how to detect these networks from time series has not been examined rigorously from a mathematical perspective. As such, a primary goal of this project was to determine a mathematically motivated procedure to identify functional networks using time series analysis. This type of analysis is useful in determining correlations between events with respect to time [15]. The measures employed in this study incorporate the effects of time delays on correlative relationships in EEG time series.

2 Methods

2.1 Participants

14 individuals (age range 10–16) with diagnoses of Asperger Syndrome or high-functioning autism spectrum disorder (ASD), and 14 healthy, typically developing (TD) age and gender-matched subjects (age range 10–17) recruited by flyer advertisements among the schools of Arlington School District, TX, participated in the study. Participants' parents were asked to evaluate their children's communicative abilities using the Pragmatic Language Observation Scale [10]. Participants scoring at or above average (90 and above) on the Pragmatic Language Observation Scale were assigned to the control (TD) group; participants with scores lower than one standard deviation below average (84 and below) were assigned to the ASD group.

The study was approved and conducted in accordance with the ethical standards of the University of Texas at Arlington Institutional Review Board, and the ethical standards prescribed in the 1964 Declaration of Helsinki and its later amendments. All parents provided their written, informed consent, and children provided written, informed assent prior to their inclusion in the study.

2.2 Electroencephalogram

Scalp EEG was recorded from 32 Ag/AgCl electrodes mounted in an electrode cap (Wavegard, ANT Inc.) with an average mastoid reference. Electrodes

were positioned according to the standard 10-20 system. A pair of bipolar electrodes were used to record vertical eye movements. Electrode impedances were maintained below 10 kΩ during recording. The EEG analog signal was digitized at a 512-Hz sample rate.

2.3 Procedures

During the EEG session, the participants were seated comfortably in a sound-attenuating booth with their eyes approximately 80 cm from a computer screen. The participants were asked to keep their eyes on the screen and to decide as accurately and as quickly as possible whether the stimulus photograph expressed fear or anger. When the face-body compound stimuli were presented, participants were told to judge the expression of the face. Stimuli were presented for 1000 ms, followed by a black screen for 2000 ms. The hand assigned to Fear/Anger response was balanced among participants. The testing started with a short training session to acquaint participants with procedures and task expectations of the experiment. Examples from all four stimulus categories were included in the training.

The study consisted of 4 blocks: 2 separate blocks when participants viewed isolated faces and bodies and 2 blocks with compound stimuli. Each block/ category consisted of 40 stimulus trials, for a total of 160 trials. The order of block presentation varied, such that half of the participants viewed a face or body-only block first, and half viewed a block with composite stimuli first. Of those who first saw a control block, half started with the isolated faces block, and half started with the isolated bodies block. For the purposes of this study, the EEG data from only one event type (anger in the body, no face shown) were used. Hereafter, this data will be referred to as "event-related" data.

2.4 Signal Processing

EEG data were analyzed using ASA 4.6 (ANT, Inc.). The continuous resting EEG data were shortened to one 500 ms long epoch and no other processing was performed. The continuous event-related EEG data were segmented into epochs of 500 ms consisting of data from 100 ms pre-stimulus onset and from 400 ms post-stimulus offset. Time points in the filtered data at which the absolute amplitude of the EEG exceeded $\pm$150 V were marked as EEG artifacts or blink artifacts. Trials containing EEG artifacts were rejected from further analyses, as were trials containing incorrect behavioral responses. Averages were baseline corrected using the 100 ms pre-stimulus portion of the epoch. Ten TD children and ten children with ASD had EEG data in both conditions satisfying the rejection criteria. Consequently, only the data from these 20 subjects were included in the analysis.

2.5 Time Series Analysis

MATLAB ® (Version 2009b, The Mathworks, Natick, MA) was used for all computations described in the sections below.

Sample cross-correlation was used to reveal correlations between signals from two electrode sites at different time lags. First, for jointly stationary time series x and y with n entries, the **cross-covariance** $\gamma_{xy}(h)$ (see [15]) at lag h is estimated by the sample cross-covariance

$$\hat{\gamma}_{xy}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(y_t - \bar{y}), \tag{1}$$

for $0 \leq h \leq n-1$. The use of the relationship $\hat{\gamma}_{xy}(h) = \hat{\gamma}_{yx}(-h)$ allows for the computation of cross-covariance for $-(n-1) \leq h < 0$. The sample **cross-correlation** $\hat{\rho}_{xy}(h)$, given by

$$\hat{\rho}_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_{xx}(0)\hat{\gamma}_{yy}(0)}}, \tag{2}$$

normalizes the cross-covariance so that $-1 \leq \hat{\rho}_{xy}(h) \leq 1$. The sign of h indicates the nature of the relationships (i.e. leading versus lagging) in the EEG data, and the magnitude of $\hat{\rho}_{xy}(h)$ indicates the strength of the correlation. The derivation of (1) assumes that the time series x and y are weakly stationary, meaning the probability distributions of any collection of their vector components with the same length is independent of time. This assumption may be reasonable for the resting condition because resting EEG time series do not exhibit trending means. While the assumption is less reasonable for the task condition, a detrending correction was implemented on the event-related time series. Even then, a violation of this assumption does not prohibit the use of this time series tool, but it may weaken its conclusions. In any case, the hypothesis of weak stationarity cannot be proved without assuming a model of the brain, and a universal model has not been agreed upon in the neuroscience community.

Each subject's event-related and resting data were divided into individual epochs of 16 data points (corresponding to 31.25 ms). Cross-correlations were computed for all pairs of nodes and all time lags h within each individual epoch.

2.6 Network Identification

Define $h_{\max}$ as the h yielding the greatest $|\hat{\rho}_{x_i x_j}(h)|$ for a given pair (i, j) of nodes. Thus, each pair of nodes and its corresponding $h_{\max}$ has an associated maximal cross-correlation $|\hat{\rho}_{x_i x_j}(h_{\max})|$. This value is stored in a matrix, S, as the entry S_{ij} in the ith row and jth column. Note that $S_{ij} = S_{ji}$, meaning S can be thought of as a weighted, undirected graph. Every vertex of this graph is connected with every other vertex, and the strength of each edge is the strength of the cross-correlation between the nodes it connects (i.e., the value of $|\hat{\rho}_{x_i x_j}(h_{\max})|$).

Since the cross-correlation defined in (2) is a function of the lag h, large entries in S need not indicate that the corresponding time series are highly similar at each point in time. In fact, supposing one time series frequently lags another (that is, the first time series looks very similar to the second after the first has been translated forward in time), the two could appear quite different at all points in time. The fact that cross-correlation allows strong but lagged linear correlations to be detected is crucial in identifying these functional connections. Nevertheless, no matter how strong these connections may be, if they only occur for large lags, one should question their significance to functional behavior because regions of the brain communicate with each other quickly [16, 17]. Thus, a correlation measure giving preference to smaller lags is desired. Fortunately, the one given by (2) has exactly this feature, since the sum in (1) includes fewer summands for larger values of h, yet the sum is always divided by n instead of $n - h$. The result is to detect functional connections over larger lags only if the relationship is particularly strong during the epoch.

2.7 Network Analysis

Once a network is established for each epoch, the relative stability of the network with respect to time was computed via edit distance. **Edit distance** between $m \times n$ matrices A and B is defined as

$$\operatorname{dist}(A, B) = \sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij} - B_{ij}|. \tag{3}$$

This computation was performed for each pair of consecutive epochs. The change from one connectivity matrix to the next matrix is called a **network transition**. Network transitions with an edit distance more than two standard deviations below the mean of the null model were considered **stable**. The null model was generated by randomizing every matrix (i.e. graph) for every subject. This randomization was executed by use of the `null_model_und_sign` function (with the default settings) in the Brain Connectivity Toolbox [14], which reassigns edge weights while preserving the weight, degree, and strength distributions of each graph. Two mean edit distances of the null model were calculated, one for each task condition. The mean for resting state network transitions was computed using both populations' randomized resting matrices, and that for event-related transitions using both populations' randomized event-related matrices. For each null model, the stability threshold was established as described above for the corresponding condition.

An alternative construction of the null models randomizes only one subject's functional connectivity matrices, creating a null model specific to each subject for each condition. This approach, however, does not accurately capture variability between subjects because individualized null models are entirely determined by the randomization of only one subject's functional connectivity matrices. Instead, the method described above offers a global perspective of typical edit distances between independent and pseudorandom networks, thereby

lending a constant threshold with which to compare the network transitions of all subjects in a given condition.

Figure 1 provides a visualization of the methods described in 2.5 through 2.7.

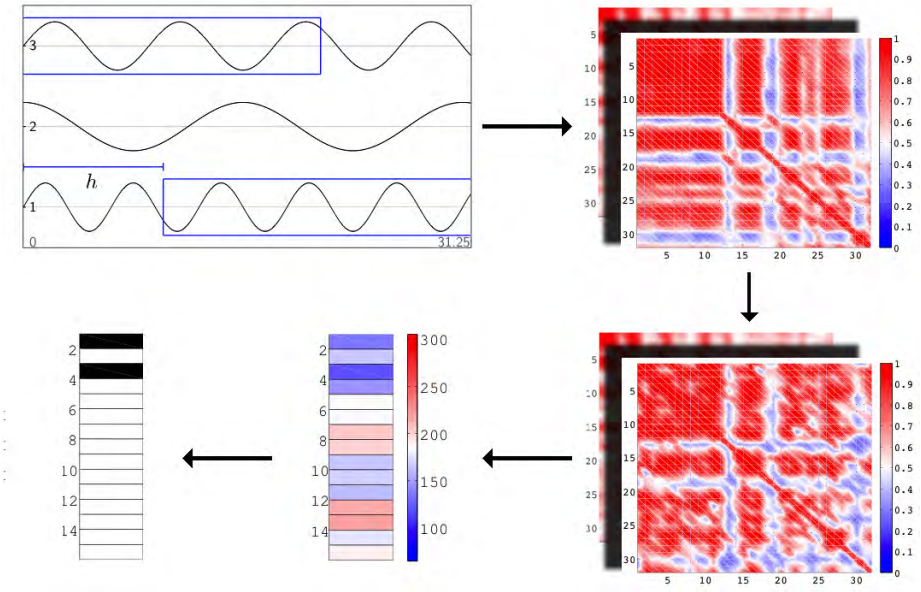


Figure 1: Each subject's processed EEG signal leads to 16 functional matrices, one for each epoch, when (2) is computed for each pair of nodes and all possible lags. For instance, the red cells seen in the top matrix indicate pairs of nodes for which there existed a lag producing a high cross-correlation value. All such matrices are randomized, and the edit distances (3) between them establishes the null model. Finally, the null model is used to detect stable network transitions between the original matrices by thresholding edit distance vectors against the two standard deviation cutoff. Stable transitions are denoted by black cells. The example shown here had stable transitions between epochs 1 and 2 and between epochs 3 and 4, as suggested by their dark blue coloring in the edit distance vector.

The use of edit distance (the ℓ-1 norm for matrix entries) is a first approach to the determination of distances between matrices (i.e., the determination of stable networks). An alternative metric, the ℓ-2 norm for matrix entries, was used in place of edit distance for an additional analysis. The **ℓ-2 norm for matrix entries** is

$$\text{dist}_2(A, B) = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}(A_{ij} - B_{ij})^2} \tag{4}$$

for $m \times n$ matrices A and B.

2.8 Network Measures

Define an **averaged stable network** as

$$\frac{1}{n}\left(S_i + S_{i+1} + \cdots + S_{i+(n-1)}\right), \tag{5}$$

where S_i is a functional matrix, n is the number of epochs in the period of quasi-stability, and i is the first epoch in that period. A **period of quasi-stability** is a series of consecutive epochs separated by stable network transitions. For example, the binarized vector shown in Figure 2 yields the following four periods of quasi-stability, each identified as a black "block":

1. The first period of quasi-stability consists of epochs 1 and 2.
2. The second includes epochs 3 through 6.
3. The third includes epochs 7 and 8.
4. The fourth includes epochs 13 through 16.

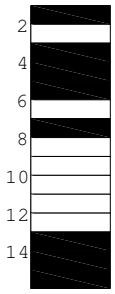
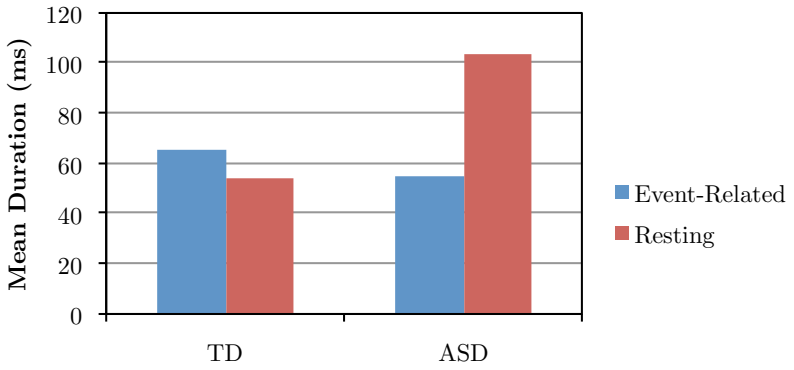


Figure 2: Binarized edit distance vector for ASD subject 110 in resting condition

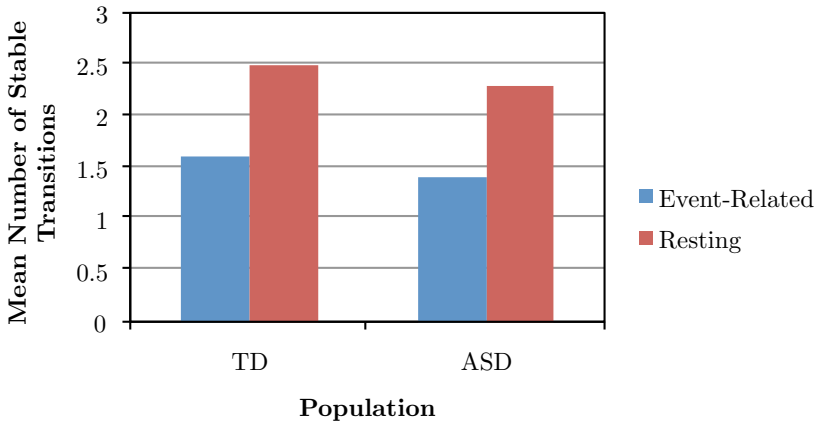
For every subject, an averaged stable network was computed for each period of quasi-stability. Several network measures were calculated for all averaged stable networks by use of the Brain Connectivity Toolbox, including diameter, radius, characteristic path length, transitivity, maximum modularity, and global efficiency [14].

3 Results

A repeated measures analysis of variance (rmANOVA) was performed for both the mean number of stable network transitions and the mean duration of stable network transitions with a between-subjects factor of population (TD versus ASD). There was an interaction effect between mean duration and population



(a) The mean durations of stable network transitions for TD children were 65.6 ms (standard deviation 29.5 ms) in the task condition and 54.3 ms (standard deviation 27.8 ms) in the resting condition and, for children with ASD, were 55.2 ms (standard deviation 38.1 ms) in the task condition and 104.2 ms (standard deviation 71.0 ms) in the resting condition.



(b) The mean numbers of stable network transitions for TD children were 1.6 (standard deviation 1.1) in the task condition and 2.5 (standard deviation 1.7) in the resting condition and, for children with ASD, were 1.4 (standard deviation 1.0) in the task condition and 2.3 (standard deviation 1.0) in the resting condition.

Figure 3: Results of null model thresholding (Note that the standard deviation values reported are not the threshold values for the null models.)

on the two conditions ($p = 0.034$) (see Figure subfig:dur). Additionally, there was a main effect of mean number on the two conditions ($p = 0.046$) (see Figure **??**). Thus, there were significant differences in the mean number of stable network transitions between the two conditions (the resting condition had a larger mean) and in the mean duration of stable network transitions between the two populations (children with ASD had longer stable networks in the resting condition).

Another rmANOVA was performed for all of the network measures for the

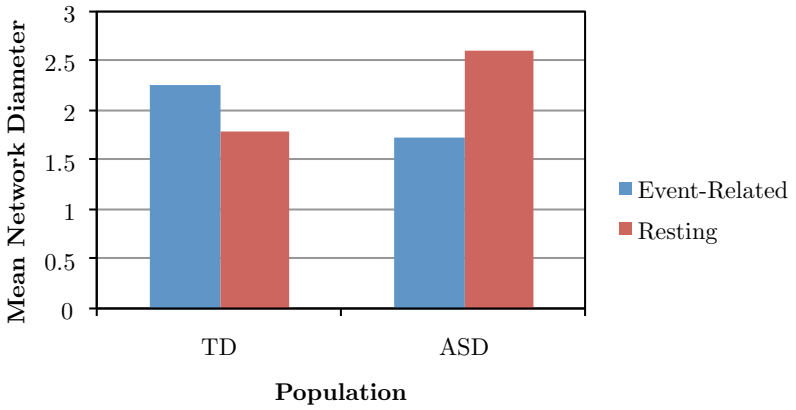


Figure 4: Result of diameter measurement of averaged stable networks. The mean network diameters of averaged stable networks for TD children were 2.26 (standard deviation 0.86) in the task condition and 1.80 (standard deviation 0.94) in the resting condition and, for children with ASD, were 1.73 (standard deviation 1.16) in the task condition and 2.61 (standard deviation 0.40) in the resting condition.

event-related and resting data with the same between-subjects factor as before. There were no main effects observed. There was an interaction effect between diameter and population on the two conditions ($p = 0.032$). Thus, there was a significant difference in the mean network diameter between the two populations (children with ASD had a larger mean network diameter in the resting condition) (see Figure 4). Additionally, there were trending interaction effects between characteristic path length and population ($p = 0.083$), as well as between radius and population ($p = 0.076$), on the two conditions.

All of the above significant results were obtained by use of the edit distance metric (3). When the ℓ-2 metric (4) was used, there were no significant results.

4 Discussion

In the neuroscience literature, there is evidence to suggest that there exists a core group of quasi-stable functional brain networks that are continually revisited while the brain is at rest [3, 8, 9]. The results from this study support the existence of quasi-stable functional brain networks, but nothing can be concluded about the number of such networks or any periodic cycling of these networks without further investigation. Future studies should attempt to classify the stable networks uncovered in this project and to assess whether or not they occur in a repetitive pattern or sequence. Moreover, further work in this area should examine the number of such networks that explain a large amount of variance of resting brain activity. Previous research suggests that four such networks explain nearly 80 percent of all resting brain activity (as can be measured

via EEG) [9].

The observation that more stable network transitions were observed during the resting condition may be explained simply. It is possible that a brain at rest cycles through the aforementioned core group of quasi-stable functional brain networks more quickly than a brain responding to a stimulus. This is because a certain functional network of the brain is activated by a response to a stimulus. Such an activation may temporarily disrupt the resting network cycle, and, consequently, it may take some time for the brain to resume its cycling after responding to a stimulus. This perturbation in the resting network cycle could explain some of the results observed in this study, specifically those concerning the number of stable network transitions.

The observation that stable networks endure for a longer period of time in children with ASD at rest is more difficult to explain. Perhaps the resting network cycle takes longer to complete in children with ASD, resulting in an increase in stable network duration. An alternative explanation is that certain networks are stable for longer periods of time due to brain abnormalities caused by ASD; however, further research is required in order to make more meaningful conclusions.

The observation that children with ASD have a higher mean network diameter may not be the best indication of functional connectivity differences. Diameter is defined as the maximum shortest path length between any two vertices where, in the context of weighted graphs, the shortest path is the smallest sum of edge weights for paths connecting the two vertices. Thus, a large mean diameter indicates an overall higher level of correlation between nodes, but this does not imply all nodes in the network are well-connected.

The nature of network transitions can be better understood upon considering the differing results of using the ℓ-1 and ℓ-2 metrics. The ℓ-2 norm is influenced more by larger entries, whereas all entries linearly contribute to the ℓ-1 norm. The fact that significances are detected in the latter but not the former suggests that the difference between stable and unstable network transitions is more related to small changes in correlation between consecutive epochs, as opposed to large changes. That is, the frequency of large changes in correlation between given pairs of nodes is more uniform across subjects and across task conditions. In contrast, from this study's results, smaller changes in cross-correlation are inferred to be more variable across the two populations and across the two conditions. Future studies should probe the effects of alternative metrics, such as the elastic net regularization, which is the sum of the ℓ-1 and ℓ-2 metrics.

Acknowledgments

We would like to thank Professor Evguenia Malaia of University of Texas at Arlington and Ph.D. candidate Jonathan Poelhuis of Indiana University for their guidance, knowledge, and support. We would also like to thank Professor Kevin M. Pilgrim of Indiana University for organizing the REU program during which this report was written, as well as Professor Elizabeth Housworth of

Indiana University for useful discussions.

Appendices

A Full Results of Null Model Thresholding

This appendix contains the results of null model thresholding for each subject in each condition. The binarized vectors in Figure 5 display which network transitions were stable; the ith entry of the vector corresponds to the transition between epoch i and epoch $i+1$.

B Selected MATLAB Code

B.1 Cross-Correlation

```
function C = cross_correlation(A)
% C = cross_correlation(A)
% Computes the cross-correlation of columns of A at all possible lags
%
% input:    A: m by n matrix
% output:   C: m by n by n array
%               C(h,j,k) is the cross_correlation of columns j and k of A
%               with lag h-1.

m = size(A,1);
n = size(A,2);
avg = mean(A,1);
AVG = repmat(avg,m,1);
auto_covariance = sum((A - AVG).^2,1)/m;

C = zeros(m,n,n);

for h = 0:m-1
    for j = 1:n
        for k = 1:n
            x = A(1+h:m,j);
            y = A(1:m-h,k);
            C(h+1,j,k) = 1/m*sum((x - avg(j)).*(y - avg(k)))/...
                sqrt(auto_covariance(j)*auto_covariance(k));
        end
    end
end
```

B.2 Network Identification

```
function [S,T] = strongest_lag(C)
% S = strongest_lag(C)
% Computes the strongest cross-correlation and associated lag from those
%   stored in C
%
```

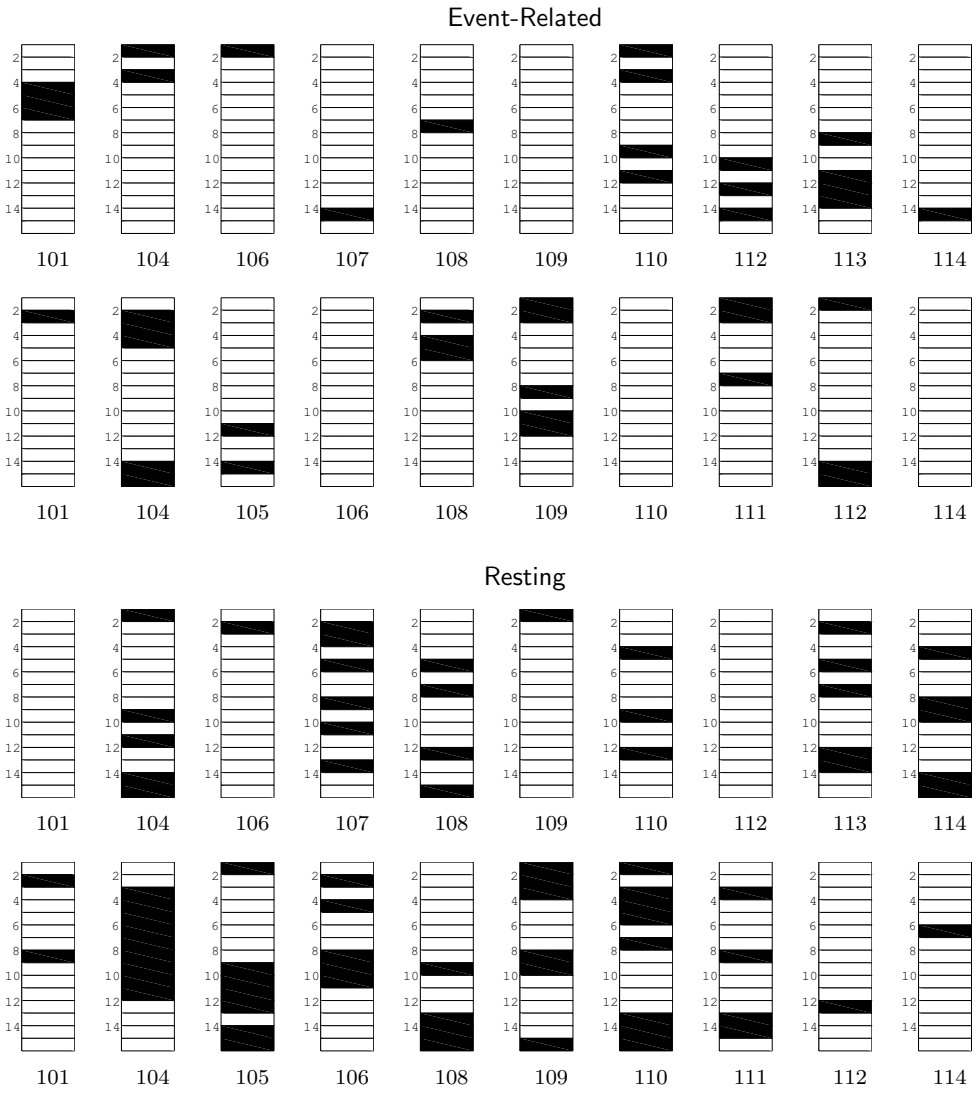


Figure 5: Individual results of null model thresholding. Stable transitions are denoted by black cells. Vectors in the first rows correspond to TD subjects, and those in the second rows to ASD subjects. Subject numbers are shown below the vectors.

```
% input:    C: m by n by n array
% outputs:  S: n by n symmetric matrix
%           T: n by n skew-symmetric matrix
%               S(i,j) is the largest entry value of
%               {abs(C(:,i,j)),abs(C(:,j,i))}, and T(i,j) is the
```

```
%               associated lag. Positive means j lags i.

m = size(C,1);
n = size(C,2);
S = zeros(n,n);
T = zeros(n,n);

for i = 1:n
    for j = i+1:n
        [S(i,j),T(i,j)] = max(abs([flipud(squeeze(C(:,i,j)));...
            squeeze(C(2:end,j,i))]));
    end
end

T = T-m*triu(ones(n,n),1) - triu(T-m*triu(ones(n,n),1),1)';
S = S + triu(S,1)';
S = S + eye(n); % Comment if self-correlation of 1 should not be shown.
```

B.3 Edit Distance

```
function D = edit_distance(C)
% D = edit_distance(C)
% Computes the edit distances between matrices of C as separated along the
%   third dimension
%
% input:    C: h by n by m array
% output:   D: h by h matrix
%               D(i,j) is the "edit distance" between C(:,:,i) and C(:,:,j)

h = size(C,3);
D = zeros(h,h);

for i = 1:h
    for j = i+1:h
        D(i,j) = sum(sum(abs(C(:,:,i)-C(:,:,j))));
        D(j,i) = D(i,j);
    end
end
```

B.4 Null Model

```
function [avg,stdev] = random_edit_distance(S,bin_swaps,wei_freq)
% [avg,stdev] = random_edit_distance(S)
% Computes the average and standard deviation of edit distance separating
% randomized versions of the matrices contained in S along the third
% dimension
```

```
%
% inputs:   S, n by n by t matrix
%           bin_swaps, average number of swaps of each edge in binary
%               randomization.
%               bin_swap=5 is the default (each edge rewired 5 times)
%               bin_swap=0 implies no binary randomization
%           wei_freq, frequency of weight sorting in weighted randomization
%               wei_freq should range between 0 and 1
%               wei_freq=1 implies that weights are resorted at each step
%                   (default in older [<2011] versions of MATLAB)
%               wei_freq=0.1 implies that weights are sorted at each 10th
%                   step (faster, default in newer versions of Matlab)
%               wei_freq=0 implies no sorting of weights
%                   (not recommended)
% outputs:  avg, the mean edit distance between randomized graphs
%           stdev, the standard deviation of those edit distances
%
% Disclaimer: part of this code is taken from null_model_und_sign.m,
% written by Mika Rubinov.

if ~exist('bin_swaps','var')
    bin_swaps=5;
end
if ~exist('wei_freq','var')
    if nargin('randperm')==1
        wei_freq=1;
    else
        wei_freq=0.1;
    end
end
S_random = zeros(size(S));

for j = 1:size(S,3)
    S_rand = null_model_und_sign(S(:,:,j),bin_swaps,wei_freq);
    S_rand(1:size(S,1)+1:size(S,1)^2) = 1; % set all diagonal entries to 1;
    S_random(:,:,j) = S_rand;
end

D = edit_distance(S_random);
D = triu(D,1);
D = D(D~=0);
avg = mean(D);
stdev = std(D);
```

B.5 Thresholding

```
function M = stable_transitions(S,avg,stdev,tol)
% M = stable_transitions(S,avg,stdev,tol)
% Identifies intervals of S along the third dimension that do not change
%   (in edit_distance) consecutively by more than avg - tol*stdev
%
% inputs:   S, n by n by t matrix
%           avg, positive number
%           stdev, positive number
%           tol, positive or negative number
% output:   M, w by 1 vector
%               The stable transitions are the entries of M.

D = edit_distance(S);
D = diag(D,1); % only consider consecutive edit distances

tol = avg - tol*stdev;
M = find(D<tol);
```

References

[1] Pablo Barttfeld, Bruno Wicker, Sebastián Cukier, Silvana Navarta, Sergio Lew, and Mariano Sigman. A big-world network in asd: dynamical connectivity analysis reflects a deficit in long-range connections and an excess of short-range connections. *Neuropsychologia*, 49(2):254–263, 2011.

[2] Danielle S Bassett and Edward T Bullmore. Human brain networks in health and disease. *Current opinion in neurology*, 22(4):340, 2009.

[3] Richard F Betzel, Molly A Erickson, Malene Abell, Brian F O'Donnell, William P Hetrick, and Olaf Sporns. Synchronization dynamics and evidence for a repertoire of network states in resting eeg. *Frontiers in computational neuroscience*, 6, 2012.

[4] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

[5] Vladimir L Cherkassky, Rajesh K Kana, Timothy A Keller, and Marcel Adam Just. Functional connectivity in a baseline resting-state network in autism. *Neuroreport*, 17:1687–1690, 2006.

[6] Yong He and Alan Evans. Graph theoretical modeling of brain connectivity. *Current opinion in neurology*, 23(4):341–350, 2010.

[7] Daniel P Kennedy and Eric Courchesne. The intrinsic functional organization of the brain is altered in autism. *Neuroimage*, 39(4):1877–1885, 2008.

[8] Thomas Koenig, Leslie Prichep, Dietrich Lehmann, Pedro Valdes Sosa, Elisabeth Braeker, Horst Kleinlogel, Robert Isenhart, and E Roy John. Millisecond by millisecond, year by year: normative eeg microstates and developmental stages. *Neuroimage*, 16(1):41–48, 2002.

[9] D Lehmann and W Skrandies. Reference-free identification of components of checkerboard-evoked multichannel potential fields. *Electroencephalography and clinical neurophysiology*, 48(6):609–621, 1980.

[10] P.L. Newcomer and D.D. Hammill. *Pragmatic Language Observation Scale (PLOS)*. Hammill Institute on Disabilities, 2009.

[11] Ernst Niedermeyer and Fernando H Lopes da Silva. *Electroencephalography: basic principles, clinical applications, and related fields*. Wolters Kluwer Health, 2005.

[12] Jurriaan M Peters, Maxime Taquet, Clemente Vega, Shafali S Jeste, Ivan Sanchez Fernandez, Jacqueline Tan, Charles A Nelson, Mustafa Sahin, and Simon K Warfield. Brain functional networks in syndromic and non-syndromic autism: a graph theoretical study of eeg connectivity. *BMC medicine*, 11(1):54, 2013.

[13] Jaap C Reijneveld, Sophie C Ponten, Henk W Berendse, and Cornelis J Stam. The application of graph theoretical analysis to complex networks in the brain. *Clinical Neurophysiology*, 118(11):2317–2331, 2007.

[14] Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

[15] Robert H Shumway and David S Stoffer. *Time series analysis and its applications: with R examples*. Springer, 2011.

[16] Olaf Sporns. *Networks of the Brain*. The MIT Press, 2011.

[17] Olaf Sporns. *Discovering the Human Connectome*. MIT Press, 2012.

[18] Cornelis J Stam and Jaap C Reijneveld. Graph theoretical analysis of complex networks in the brain. *Nonlinear biomedical physics*, 1(1):3, 2007.

[19] LQ Uddin, K Superkar, C Lynch, A Khouzam, J Phillips, C Feinstein, et al. Salience network based classification and prediction of symptom severity in children with autism. *JAMA Psychiatry*, 10, 2013.

Classifying the Structure of 8 Column Sets in Hadamard Matrices of Size 24

Wade Bloomquist

Abstract

It is known that there exist 60 equivalence classes of matrices of size 24. One can relate 59 of these equivalence classes through an operation known as switching. This switching can be thought of as being based on weight 4 codewords in the code generated by the matrix. The Golay code is generated by two of of these equivalence classes, one of which is the equivalence class not represented before. This code does not admit weight 4 codewords so we hope to extend this idea to weight 8 codewords.

1 Introduction

The Golay code is one which has come up in many contexts throughout mathematics. In geometry the Golay code can be used to construct the Leech lattice which is used to find a current record holding sphere packing. In algebra the Golay code was used to build the Conway group which is instrumental in the construction of the monster group to finish the classification of finite simple groups. These examples provide a strong motivation in trying to look at the code to help in other problems.

Specifically, it is not known how all 60 equivalence classes of Hadamard matrices of size 24 can be related. In fact 59 have already been related but there is one that is left out. However, this final matrix and one already understood both generate the Golay code. We see that the relationships already known are based on 4 column structures, but the Golay code does not allow this as a possibility for matrices that generate it. Instead we try to use 8 column sets to find similar structure. We use three methods. The first involves defining an invariant on 4 column sets and taking this invariant on all 4 column sets taken from the 8 column sets. Next a canonical form was defined on 8 column sets. Finally we look at the automorphism group and how its action on 8 column sets is structured.

This project was done with help of the GAP software [4]. We also made use of the GUAVA and GRAPE packages. This has allowed for observations that will be included in the paper without proof.

2 Background

2.1 Coding Theory

Algebraic coding theory will provide the setting in which this project is framed. We will thus need to introduce some vocabulary primarily based on that presented in [5].

Definition 2.1 A *code* is a set of strings, called codewords, that are built from elements of a set, called a library, say K.

A code can be thought of as a way of sending information across a channel that may contain noise which will alter the intended message.

Definition 2.2 A *k-error correcting code* is one in which k errors can occur, and the original message will still be recovered.

If we allow K to be a field then we can define a code over K to be a subset of K^n, namely all of the n-tuples over the field K. In this situation K provides a library for possible entries in a message to be sent. Most commonly this field K is taken to be $\mathbb{F}_q$, the finite field over q elements. This project deals with linear binary codes.

Definition 2.3 Let $\mathcal{C}$ be a code. We define $\mathcal{C}$ to be *binary* if the library that the code is built from only has two elements.

The most commonly used example of this will be taking $K = \mathbb{F}_2$. Then we see that our code is a subset of $\mathbb{F}_2^n$.

Definition 2.4 Let $\mathcal{C}$ be a code. We define $\mathcal{C}$ to be *linear* if the sum of any two codewords in $\mathcal{C}$ is another codeword of $\mathcal{C}$.

We can then observe that a linear code over $\mathbb{F}_2$ is a vector subspace of $\mathbb{F}_2^n$, for some positive integer n.

Definition 2.5 Let $\mathcal{C}$ be a linear code. Then define a *generator matrix* of $\mathcal{C}$ to be a matrix whose rows are the basis of $\mathcal{C}$.

Definition 2.6 Let $\mathcal{C}$ be a code over $\mathbb{F}_q$. Then the *dual code* to $\mathcal{C}$, denoted $\mathcal{C}^\perp$, is defined by

$$\mathcal{C}^\perp = \{x \in \mathbb{F}_q^n | x \cdot c = 0 \quad \text{for all c in} \quad \mathcal{C}\}.$$

We call $\mathcal{C}$ *self dual* if $\mathcal{C} = \mathcal{C}^\perp$

Definition 2.7 Let $\mathcal{C}$ be a code and let c be a codeword of $\mathcal{C}$. Then the *weight* of c is defined to be the number of nonzero entries in c.

Definition 2.8 Let $\mathcal{C}$ be a code. We say that $\mathcal{C}$ is *doubly even* if every codeword in $\mathcal{C}$ has a weight that is a multiple of 4.

We now provide a list of terms about codewords that will be used throughout the paper

1. A duad is a 2 entry structure.
2. A tetrad is a 4 entry structure.
3. A sextet is a 6 entry structure.
4. An octad is an 8 entry structure.
5. A dodecad is a 12 entry structure.

Definition 2.9 Let $\mathcal{C}$ be a code with generating matrix A and c a codeword. Then we define a *column set* of c to be the set of columns that are indexed to match the indexed nonzero entries of c.

As a technical detail, when a code is generated by a Hadamard matrix, as will be discussed later, we actually change the matrix slightly before generating the code. We however will take column sets to be from the original Hadamard matrix.

2.2 The Golay Code

The Golay code, more technically the extended binary Golay code, is a code on 24 elements over $\mathbb{F}_2$. The structure of this code allows for the detection of any 7 errors or the correction of any 3 errors. We also note that the Golay code has minimum weight 8, is self dual and is doubly even. The automorphism group of the Golay Code is the Mathieu Group on 24 elements.

Definition 2.10 Let $\mathcal{C}$ be a binary linear code. Then an *automorphism* of $\mathcal{C}$ is defined to be a permutation that leaves the codewords invariant.

Proposition 2.11 *Any two distinct octads in the Golay code can only overlap in 0, 2 or 4 ways.*

Proof We begin by noting that the sum of two codewords is a codeword in the Golay code. Thus we can take any overlapping octads and find a new codeword that is of weight $16 - 2l$, where l is the number of positions that overlap. This tells us that an odd number of overlapping positions gives a codeword of weight not a multiple of 4. This cannot happen as the Golay code is doubly even. We know the overlap cannot be all eight entries as we have chosen them to be unique. Now we see that we cannot not have 6 overlapping entries because this would imply a codeword of weight 4, but the minimum weight of the Golay code is 8. Thus we are left with 0, 2 and 4 as possible amounts of overlap between two octads. □

Proposition 2.12 *A 5 column set determines a unique 8 column set.*

Proof Assume that the 5 column set is in two different 8 column sets. Thus two weight 8 codewords overlap in 5 entries. As we are in a linear code the difference between codewords is also a codeword. This would give a weight 6 codeword, which is a contradiction. Thus A 5 column set is determined to be in a unique 8 column set. □
Now we step back and want to look at 4 column sets. These are of even more interest to us due to our invariant being defined on them.

Proposition 2.13 *We see that any 4 column set belongs to exactly five 8 column sets.*

Proof We begin with a 4 columns set, say A. If we add another column to A we have a 5 column set which uniquely determines an octad, say B. We will also arrive at B if we take any column in B that was not in A to begin with. We see that this allows us to associate 4 columns that can be added to A to give an octad. This follows from the above proposition as we cannot have octads intersect in more than the 4 entries in which they are forced to intersect in since we began with A. As we have 24 total columns and 4 are used to make A, we have 20 possible columns to add to A. This tells us that 5 octads can be made as $\frac{20}{4} = 5$. □

2.3 The Mathieu Group

The definitions in this section are taken from [1]

Definition 2.14 Let G be a group and X a set. We then define a *group action*, of G on X, to be a function by

$$G \times X \to X, \quad (g, x) \mapsto g \cdot x$$

where $(gh) \cdot x = g \cdot (h \cdot x)$ for all g,h in G and all x in X. We also require $e \cdot x = x$ for all x in X.

This can be described more intuitively by looking at G as the symmetries of the set X where the action of an element of g is a way of permuting trough the symmetries.

Definition 2.15 Let G be a group and X a set. We call the action of G on X *transitive* if any element in the set X can be sent to any other element in the set through an action of an element in G.

Viewing this concept from the viewpoint of the orbit structure of the action is more enlightening.

Definition 2.16 Let G be a group, X a set, and x and element of X. Then we define the *orbit* of x, O_x, to be the set of all points x can be mapped to through the action of an element of G.

This allows us to use an alternate, but equivalent, definition of a transitive action.

Definition 2.17 Let G be a group and X a set. We call the action of G on X transitive if there is only one orbit.

Now we look to generalize the notion of transitivity to multiply transitive actions.

Definition 2.18 Let G be a group and X a set. We say the action of G on X is *k-transitive*, for k a positive integer, if any k-tuple of distinct elements of X can be mapped to any other k-tuple of distinct elements of X.

Multiple transitivity is a very rare property outside of the symmetric and alternating groups. Mathieu was very interested in exploring examples of these groups at the end of the 1800's. This led to the construction of the Mathieu group, M_{24}. We note that M_{24} acts 5-transitively on a set of 24 element and is a simple sporadic group.

2.4 Hadamard Codes

We now begin to look at the specific codes that we will be using during our project. These are codes that are generated by Hadamard matrices.

Definition 2.19 A *Hadamard matrix* of size n is a $\{+1,-1\}$ square matrix that has mutually orthogonal rows. Equivalently we see this implies that $HH^T = nI$.

These matrices, when they exist, will be the maximal determinant matrices of a given size.

Proposition 2.20 *Hadamard matrices attain the Hadamard bound,* $\det(H) \leq n^{\frac{n}{2}}$.

Proof We can see that this is a bound by looking at the determinant as the volume of the parallelepiped spanned by the rows. This gives us that the determinant of a matrix will be less than the product of the magnitudes of the rows. This then immediately tells us we reach this maximum when the product of the magnitudes is maximized, namely when the we have only orthogonal rows. □

We see that size 1 is trivial and size two is simply a row of ones and a row with a one and a negative one.

We observe that Hadamard matrices are not known to exist in all cases. What is known is that there are Hadamard matrices of size 1, size 2, and any other size must be a multiple of 4. This is discussed further in [3].

Definition 2.21 Let A and B be two matrices. We call them *Hadamard equivalent* if a series of row permutation, row negations, column permutations, and column negations, can map A to B.

It has been seen that there are precisely 60 different equivalence classes of Hadamard matrices of size 24 [7, 6] An operation known as switching allows us to relate 59 of these equivalence classes. We refer the reader to figures 1 and 2

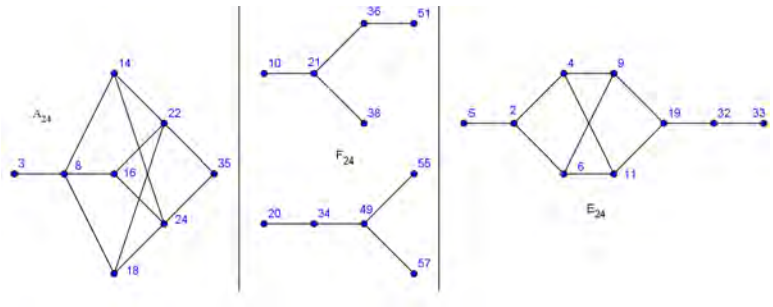


Figure 1: Diagram showing relations through switching

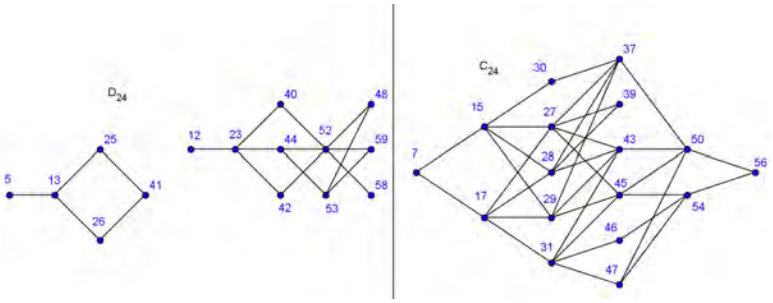


Figure 2: Diagram showing relations through switching

where this mapping is shown explicitly, where the notation of how the graphs are broken up per code is the same as [2].

It should be noted that while the 5 different separations of the graphs may seem distinct, if we introduce the transpose operation we are able to fully complete the mapping. This includes taking This includes mapping matrix 33 to matrix 58 which will be discussed more later. The 60^{th}, matrix found through the Paley construction, is left out. Our central motivation that pushes us forward will be to find some structure that allows us to relate these matrices. First we will look at the switching operation in hopes of gaining some insight. Switching operations are important because although they will not necessarily keep a matrix in its same equivalence class it will return another Hadamard matrix. We look at an $n \times n$ Hadamard matrix that is normalized in a way that the first four columns are as the transpose of the following

$$\begin{pmatrix} 1 & ... & 1 & - & ... & - & - & ... & - & 1 & ... & 1 \\ 1 & ... & 1 & - & ... & - & 1 & ... & 1 & - & ... & - \\ 1 & ... & 1 & 1 & ... & 1 & - & ... & - & - & ... & - \\ 1 & ... & 1 & 1 & ... & 1 & 1 & ... & 1 & 1 & ... & 1 \end{pmatrix}$$

.

Where the remaining columns are being ignored for now. Then we call a switching operation the process of replacing the $4 \times \frac{n}{4}$ block of ones that start this structure with negative ones. This the first 4 rows of our new matrix will be the transpose of the following

$$\begin{pmatrix} - & ... & - & - & ... & - & - & ... & - & 1 & ... & 1 \\ - & ... & - & - & ... & - & 1 & ... & 1 & - & ... & - \\ - & ... & - & 1 & ... & 1 & - & ... & - & - & ... & - \\ - & ... & - & 1 & ... & 1 & 1 & ... & 1 & 1 & ... & 1 \end{pmatrix}$$

.

Now we look at a specific case of Hadamard matrices of size 24 and how switching is seen then. We will be looking at the Kronecker product matrix, S in our diagram, $H_2 \otimes H_{12}$ and a slight generalization of it. The Kronecker product matrix can be written as a block matrix of the form

$$S = \begin{pmatrix} H_{12} & H_{12} \\ H_{12} & -H_{12} \end{pmatrix}$$

We now instead want to look at a slight generalization that will allow for a better understanding of the switching operation. Suppose we have

$$\mathcal{H} = \begin{pmatrix} A & B \\ A & -B \end{pmatrix}$$

Now we call a_1 the first row of A, a_2 the second row of A, b_1 the first row of B, and b_2 the second row of B. Now if we permute b_1 and b_2 to make a new matrix B^* we see that for

$$\mathcal{H}^* = \begin{pmatrix} A & B^* \\ A & -B^* \end{pmatrix}$$

we have that $\mathcal{H}$ is related to $\mathcal{H}^*$ through the switching operation defined above.

Now we must discuss how to go from a Hadamard matrix to a code. This is done through a simple algorithm

1. We start with a Hadamard matrix and begin by normalizing the first row to entirely negative 1.
2. Next we replace every element, x, with $log_{-1}(x)$. Explicitly this involves replacing 1 with 0 and -1 with 1.
3. Now we have a generating matrix for our code. So we are able to take the span of the rows of this new matrix over $\mathbb{F}_2$ to build a code.

It is known that two matrices produce the Golay code discussed above. One of them is matrix 58 mentioned previously. The other is the matrix that does not fit with the other 59, the Paley matrix, which is matrix 60. This leads us to believe that due to the special nature of the Golay code it will be possible to find some structure in the matrices that will relate them.

3 Results

In a way this switching operation has just used 4 column sets to relate equivalence classes of Hadamard matrices. This seems very logical as the codes that are generated by these matrices all have minimum weight 4. The Golay code has minimum weight 8, we can see if we do have a matrix that the 4×4 block can be made on this implies a weight 4 codeword. Thus we are unable to discuss matrix 58 and matrix 60 in terms of switching. This lead us to pursue an understanding of these 8 column sets in hopes of finding structure that would allow for relationships to be built.

3.1 Invariants

In order to begin our examination of the structure of these matrices we must first find an invariant. This will be important for our discussion because it will allow us to look at entire equivalence classes of matrices rather than a single representative. This means we must find qualities of these matrices that will be unaffected by the operations of Hadamard equivalence The invariant that we will be using for our investigation is called the type.

Definition 3.1 Let m be the number of rows which contain an even number of ones in an $n \times 4$ matrix. Then we define *type* to be

$$\text{type} = min(\frac{m}{4}, |\frac{n-m}{4}|)$$

.

In our case where the size of the matrix is 24×24 we see that the type must be between 0 and 3. We also note that the switching operation is acting on 4 column sets where the type is 0.

Proposition 3.2 *The type of matrix is an invariant to equivalence classes of matrices up to Hadamard equivalence.*

Proof We first observe that permutations of rows and columns will not change the type as by definition the order of these do not matter. Now we look at a row negation. This will take a row with an even number of positives to a row with an even number of negatives. The size of our row is even so this means we still have an even number of positives. Similarly a row with an odd number of ones is sent to a row with an odd number of ones. Now we look at column negations. We see that an alternative way of describing the type would be as the minimum of the number of rows with an even number of positives divided by 4 and the number of rows with an odd number of positives divided by 4. What a column negation does is send rows that had an even number of positives to a row that has an odd number of positives. As this minimum is taken at the end we really are not changing the value. □

Definition 3.3 Let A be a $n \times k$ matrix. Define the 4-profile of A to be the break down of the $\binom{k}{4}$ ways of taking a 4 column set from the k column set as vector, v, with components corresponding to the number of combinations with each type. Specifically we see that v_i is the number of combinations with type of i.

Proposition 3.4 *If an 8 column set is partitioned into two 4-sets, then the 4 sets will have the same type.*

Proof We take an 8 column set then we see that it is Hadamard equivalent to an set of 8 columns when an even number of positive ones in each row. This follows immediately form the Golay code being a self dual code. Thus if we take two 4 column sets they will be forced to divide the even number of positive ones columns. Thus we have that the 4 columns will have the same type. □

In our case this will provide a 4-vector such that the first entry is the number of combinations that have type 0, the second is the number of combinations of type 1, and so on. We have three results concerning the 4-profile breakdown of 8 column sets of Hadamard matrices.

Theorem 3.5 *First we see that for the Paley matrix all 759 support column sets have the 4-profile* $(0, 0, 30, 40)$.

Theorem 3.6 *Matrix 58 has two distinct 4-profiles, namely* $(0, 0, 30, 40)$ *and* $(0, 4, 14, 52)$. *Where 264 support column sets have the first 4-profile and 495 have the second.*

This is less of a full proof and more of current observations that have helped us toward an answer so far

We begin by observing through the construction of G above we can view it as

$$G = \begin{pmatrix} A & B \\ A & -B \end{pmatrix}$$

where A and B are Hadamard matrices. We first observe that we can normalize G to have a first row that is entirely -1. This makes the first row of A entirely -1 and the first row of B entirely -1. Then from the structure of G in terms of A and B we see that the thirteenth row of G has -1 for the first 12 entries and positive 1 for the last 12. Thus we can conclude that in the code generated by G we have the codeword whose first 12 entries are one and last 12 entries are 0 and the compliment whose first 12 entries are zero and last 12 entries are 1.

Now we look to the different ways in which an octad can intersect a dodecad. We base our discussion on the observation that if an octad, A, overlaps with a dodecad, B, in n places, then a codeword of weight $20 - 2n$ will be produced. This can be seen as we are taking 20 entries and any overlap removes one nonzero entry from both A and B when they are summed. This tells us that any odd overlap is immediately ruled out as it would imply a codeword not divisible by 4 and the Golay code is doubly even. Next we see that if there are 8 overlapped entries we find a weight 4 codeword, which is not possible as the Golay code has minimum weight 8. Similarly, if there is no overlap then we would have a weight 20 codeword, and using that the Golay code contains the codeword of weight 24 the compliment of this codeword would be weight 4. Thus the only possibilities are for the octad to intersect 2,4, or 6 times with the dodecad.

Now we observe that there are $\binom{12}{5} = 792$ ways in which we can take a 5 column set from the first 12 columns. We know from above that each of these 5 column sets will determine an octad that intersects these first 12 columns in 6 positions. Thus we see that there are $\frac{792}{6} = 132$ octads that intersect the first 12 columns in 6 positions. Following exactly we have 132 octads that intersect the last 12 columns in 6 positions. Thus we have 264 octads which can be characterized in a way in which they have 2 identified columns

We see that if we take our tetrad to have neither special column we will have $\binom{6}{2} = 15$ type 2 tetrads. Similarly if we have both special columns we will have $\binom{6}{4} = 15$ type 2 tetrads. Now if we take one special column but exclude the other we will have $2 * \binom{6}{3} = 40$ type 3 tetrads. Next we see that if we look to the octads that split evenly among the first and last 12 columns we can choose the 4 columns that are overlapped in $\binom{12}{4} = 495$ ways. Now from our discussion above we know that a tetrad can be used to uniquely determine a sextet. This means if we take a tetrad then we can extend it to a sextet that either has overlap with the first 12 columns are we are in the first case mentioned or does not in which case either have overlap with the first 12 columns in our remaining 2 columns, and we are in the first case, or we do not and we are in a 4-4 split situation. Thus we only have the type 4-4 splits that have been determined already.

Theorem 3.7 *The 8 matrices found through switching operations on $H_2 \otimes H_{12}$ all have 4-profiles of $(6, 0, 8, 56)$. In each of these codes there are exactly 495 weight 8 codewords.*

Proof This observation is concerned primarily on why these 4-profiles contain six type 0 tetrads. Take one of these matrices, say A. Then we know that

$$A = \begin{pmatrix} H & H \\ H' & -H' \end{pmatrix}$$

for some H and H', which are Hadamard matrices. We then see that if a tetrad is taken to be split evenly between the first 12 and last 12 columns we will have a type 0 tetrad. This allows us to look instead for the number of ways in which we can partition one of these tetrads. This is simply $\frac{1}{2} * \binom{4}{2} = 3$. Then we see that each of these partitions corresponds to 2 type 0 tetrads. Thus we have 6 type 0 tetrads. We also see that we have a total of 495 octads by using the fact that octads can be obtained by gluing tetrads. Then we simply need to take the number of ways we can build a tetrad form the first 12 columns and glue them to the necessary columns in the last 12. So we have $\binom{12}{4} = 495$ tetrads. □

4 Canonical Forms

We determine a canonical form that help us to determine how many truly unique 8 column sets we have.

Definition 4.1 Let A be an $8 \times n$ matrix. We define a canonical form of A to be the lexicographically minimal matrix in the equivalence class that A belongs to.

The use of a lexicographic ordering is arbitrary. Any system of ordering that would allow for a minimal matrix to be chosen may be used.

Proposition 4.2 *For the Paley matrix there was found to be a single canonical form. Matrix 58 was found to have two canonical forms, one corresponding to each 4-profile.*

Also, it should be noted that the canonical form for the Paley matrix and the canonical form matrix 58 that correspond to the $(0, 0, 30, 40)$ 4-profile do not match each other.

These propositions were found through an exhaustive computational search using GAP. This process followed these steps

1. A list was made out of all possible permutations of the 8 column set.
2. Each permutation was normalized so that the first column is entirely negative.
3. Each normalized permutation was checked to see if it were minimal to the current least element under each possibility for a row normalization that gave an entirely negative row, which was sorted so the entirely negative row was on top.
4. The overall minimal element was the canonical form.

4.1 Automorphism Group

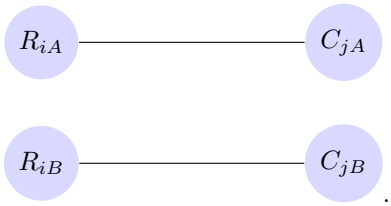
We now use the automorphism group of the Golay code, M_{24}, to try and determine more structure within the generating matrices. First we recall the definition of the automorphism group of a matrix.

Definition 4.3 The automorphism group of a {+1,−1 matrix is the group of pairs, (P, Q), of matrices such that $P^{-1}HQ = H$ and P and Q are signed permutation matrices.

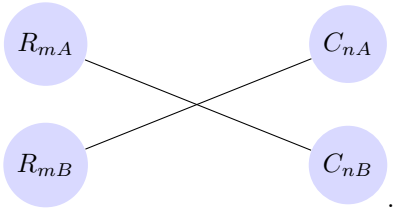
A problem arises as to how we can find these automorphism groups to help study this action better. We use a solution suggested by B. McKay that allows us to exploit existing graph-isomorphism software [8, 9].

Definition 4.4 A McKay graph is a bipartite graph that represents a plus one-minus one matrix through the following correspondence: first each row is given two vertices on the left and each column two vertices on the right. Each pair of vertices corresponding to a row is connected by two edges to each pair corresponding to a column Then when looking at a specific entry in the matrix a positive number leads to uncrossed edges between the two row vertices and the two column vertices.

As an example let M be a matrix. Then in the McKay graph of M if M_{ij}=1, then



We then look at say $M_{mn} = -1$, then



We note that we also color the rows and the columns to distinguish them as being separate This is a manifestation of not allowing transposes in our definition of Hadamard equivalence.

This allows to use the Nauty program, included in the GRAPE package, within GAP. Hadamard equivalence would be a very difficult problem to check due to the possibilities of any row or column permutation as well as negations. Although graph isomorphism is a difficult problem, it has been well studied. This allows us to instead check whether the McKay graphs are isomorphic as a

test of equivalence. More significant to the discussion at hand is the ability to find automorphism groups. The automorphism group of a graph is something that can be found with relative ease.

We currently have not worked out all of the details on when it is possible to make an isomorphism between the automorphism group of the McKay graph and the automorphism group of a matrix. This is known not to work if rows of the matrix are duplicated or are negations of other rows. We believe this isomorphism to hold when these situations are avoided. Under this assumption we move forward.

We look to see how the orbit of the automorphism group acting on these 8 column sets behaves. We begin with the Paley matrix.

Proposition 4.5 *The automorphism group of the canonical matrix from the 8 column sets in the Paley matrix is of size 96 and can be represented as* $GL(2,3) : C_2$. *We also see that the 8 column sets fall into a single orbit of size 759.*

The other matrix has a structure that is not understood as well. We see that there are two automorphism groups to be discussed. One for each canonical matrix found.

Proposition 4.6 *The automorphism group corresponding to the canonical form which matches to the* $(0, 0, 30, 40)$ *profile in matrix 58 is of size 960 and can be represented as* $C_2 \times (A_5 : C_4)) : C_2$. *There is a single orbit that corresponds to this automorphism group.*

Proposition 4.7 *The automorphism group corresponding to the 4-profile* $(0, 4, 14, 52)$ *has an automorphism group of size 256 which can be represented as* $C_2 \times (((C_4 \times C_2) : C_2) : C_2) : C_2)) : C_2$. *This is split into two orbits one of size 165 and the other of size 330.*

References

[1] N.L Biggs and A.T White: Permutation Groups and Combinatorial Structures. London Mathematical Society Lecture Note Series, 33. Cambridge University Press, Cambridge, 1979.

[2] Conway & Pless. On the Enumeration of Self-Dual Codes, Journal of Combinatorial Theory Series A 28 (1980) 26-53.

[3] J.H. Conway and N.J.A Sloane: Sphere Packings, Lattices and Groups.A Series of Comprehensive Studies in Mathematics, 280. Third Edition. Springer, New York, 1999.

[4] The GAP Group, GAP – Groups, Algorithms, and Programming, `Version 4.6.5; 2013. (http://www.gap-system.org)`

[5] Human & Pless. Fundamentals of Error-Correcting Codes, Cambridge University Press, New York, 2003.

[6] N. Ito, J. S. Leon, and J. Q. Longyear. Classification of 3 (24,12,5) Designs and 24-Dimensional Hadamard Matrices. Journal of combinatorial Theory, Series A, 31. pg. 66–93. 1981.

[7] H. Kimura. New Hadamard Matrix of Order 24. Graphs and Combinatorics, Vol 5. Issue 1. pg. 235–242. 1989.

[8] B. McKay. Hadamard Equivalence via Graph Isomorphism. Discrete Mathematics, 27, pg. 213–214. 1979.

[9] B. McKay. Practical Graph Isomorphism. Proceedings of the Tenth Manitoba Conference on Numerical Mathematics and Computing, Vol. I. Congressus Numerantium, 30. 1981.

[10] W. Orrick: Switching Operations for Hadamard Matrices. SIAM Journal on Discrete Mathematics, 22, pg. 31–50. 2008.

[11] N. J. A. Sloane A Library of Hadamard Matrices, available online at `http://www.research.att.com/~njas/hadamard/ (2008)`.

Determining Signal to Noise Ratios as Precursor to Determining Order Parameters in Light Microscopy Images of Microtubule Arrays

Allison Brumfield

Abstract

The microtubule polymers inside plant cells form various 2-dimensional patterns with functional implications for cell growth. The amount of noise in microtubule images poses a challenge for any algorithm in determining order parameters of polymer arrays from images by lowering the ability to distinguish microtubules in live-cell microscopy images. We develop a method by which to test a 2-Component Mixed Gaussian technique for calculating the signal to noise ratio as a descriptor of image quality and microtubule distinction. We develop a simulator in MATLAB to generate images with known polymer array order and signal to noise ratio. We define and determine the validity of this method for determining signal to noise ratios to account for the camera noise and the low spatial frequency noise associated with background fluorescence in live-cell microscopy images.

1 Biological Introduction

1.1 Microtubule Cytoskeletal Arrays in Plant Cells

Cellular morphogenesis or the biological process that results in a change in cell shape is a mechanism of plants for growth. The process of morphogenesis is accomplished by organizational changes in the microtubule cytoskeletal array [4]. Microtubules are tubular polymers found in eukaryotic cells that are 24 nanometers in diameter and can be 10s of microns long. The microtubules exist along the interior of the cell at the plasma membrane and form a scaffolding for the cell wall. During processes such as cellular morphogenesis, plant cells do not retain the same microtubule structure.The microtubules are not tied to a fixed, central point and can move around the cell wall forming different cytoskeletal arrays, or the patterns from microtubule arrangements. Some of these arrangements have a high degree of order and others appear completely random.

1.2 Shaw Lab and Light Microscopy

Shaw Lab is interested in plant cell morphogenesis and the organization of cytoskeletal arrays. Using florescence microscopy, single frame and time-lapse images can be collected. By altering the plant cells genes, the microtubules will fluoresce under specific wavelengths of light. The photons emitted are captured by the camera, which provides a visual of the cytoskeletal organization within an individual cell. This series of time-lapse images allow biologists, including Shaw Lab, to study the dynamics of cellular morphogenesis and microtubule cytoskeletal array organization.

1.3 Current Research

Much has been determined about the nature of microtubules. For example, the microtubules do not move in a traditional sense where the entire length of a polymer moves as a unit. By using a laser to " turn off " the florescence in a small strip of the polymers, it can be observed that the polymer lacking florescence did not move but was stationary and eventually disappeared. This series of imaging led to the conclusion that microtubules move by adding new polymer to one end or polymerizing and subtracting polymer from the other or depolymerizing.

Researchers have determined how the microtubules accomplish movement, but they are currently seeking to understand why microtubules form specific cytoskeletal arrangements with certain organizational properties and the transitions that occurs between the ordered and disordered stages. During morphogenesis, the microtubules originate in a seemingly disordered arrangement and move to a more ordered state. A pattern has been determined that organization occurs first in the center of the cell and then expands outwards towards the poles. At certain stages not only is there a transition from disorder to order, but also a non-uniform distribution of order within a single cell.

The source of one challenge from working with cytoskeletal array organization is that individuals may have different definitions of the boundary between ordered or disordered. Individual's opinions cannot be compared in such a way as to conclude something rigorously. Currently cells must be described and classified by hand according to their order and organization. Differences of opinion at this stage may lead to inconclusive results. Ultimately, there needs to be a rigorous technique to extract the order parameters or quantify order.

2 Goals

First, in order to quantify order, it must be defined in the context of microtubule arrays.

Second, the accuracy of the technique used to calculate the signal to noise ratio must be determined. In order to define the order of a given cell based only a single image, the properties of that image need to be determined which will shape the confidence and trust placed in the resulting quantified order.

Unfortunately, calculating signal to noise properties can be difficult with images of microtubules and the validity of the calculation must first be established.

3 Method

3.1 Definition of Order in Microtubule Cytoskeletal Arrays

3.1.1 Order Parameters in Physics

As a basis of framing the order and organization of microtubule cytoskeletal arrays, we consider the comparison of microtubules to liquid crystal materials. Liquid crystal materials are those that exist in a unique phase which is a hybrid state that exhibits properties of both solids and liquids [1]. In many ways liquid crystals are comparable to microtubules since both have rod-like structures (long and narrow with one axis that is significantly longer than its orthogonal axis) and change between phases having different order and organization. Such phases include the crystalline or solid state, which has extensive order in all three spatial dimensions, the mesogenic or liquid crystal state, which has orientation tendencies in one dimension, and the isotropic or liquid phase that has no order in any direction [5]. Similarly in cellular morphogenesis, the microtubule array transitions through stages starting with no order or organization to slight organization and finally arrives with an arrangement that is ordered in all dimensions assuming that microtubule arrays are planar.

Of particular interest are the liquid crystal phases where there is only partial order and the techniques of attempting to describe that order. There are several classifications of liquid crystal phases which differ in their arrangement and construction of order. The nematic phase has no positional order but has a general directional orientation such that the arrangement is ordered in one dimension and disordered in the remaining two dimensions [5]. Focusing on only the one ordered dimension, the degree of order is often described by the order parameter

$$S = \frac{1}{2} < 3\cos\theta^2 - 1 >$$

where θ = is the angle by which the long axis of the liquid crystal is rotated from the horizontal axis of an external reference axis, which describes how much variance is in the distribution of orientation angles [1]. The result is a score of alignment of the liquid crystals.

The other phase of interest is the smectic phase, which has positional order in 1 or 2 dimensions such that the liquid crystals are organized into strips or planes. The smectic phase order is a descriptor of how the material is grouped and arranged [5].

While the description of order can be roughly translated to the microtubule case, the comparison is not perfect. One biological property of microtubules is that they often have curvature along the length of the polymer, but liquid

crystals are perfectly straight. As a result the nematic order parameter S cannot be directly applied to microtubule arrays.

3.1.2 Microtubule Definition of Order

We define order parameters for use in describing the organizational structure of microtubule arrays.

Definition 3.1 The dominant angle of orientation, α, is the average angle from the horizontal axis of the external reference axes imposed on the cell.

Definition 3.2 The vector through the origin with angle given by the angle or orientation is the director vector, $\vec{n}$.

Definition 3.3 The co-alignment spacing, s, is the average spacing between microtubules orthogonal to $\vec{n}$.

Definition 3.4 The alignment or overlap spacing, t, is the average spacing between microtubules along $\vec{n}$. It is reasonable for $t < 0$ if the microtubules overlap each other.

The angle of orientation describes the rotational structure and is the description of the nematic order in the microtubule array. Together the co-alignment and alignment spacing describe the positional arrangement and can be considered a method of describing the smectic order in microtubule arrays.

3.2 Image Simulation

In order to realistically simulate images of microtubules and microtubule arrangements, create a tool capable of calculating the precise signal to noise ratio, and generate images with known order parameters and organization, we developed a program in MATLAB that constructs images by simulating each major factor that contributes to real florescence microscopy images. The aspects of the imaging process considered were the biological distribution of microtubules and several different aspects of noise including: shot noise, diffraction, and digital noise. We designed a user interface, shown in Figure 1, that allowed the user to vary over 20 parameters and parameter distributions and view the resulting image enabling different parameter spaces to be examined.

3.2.1 Biological Distribution

The biological distribution can be divided into three aspects: density and related parameters, curvature, and rotational and translational order. Using these parameters the majority of biological properties of microtubules can be observed.

Density, the amount of polymer per area, when considered as an area 1 pixel wide can be considered total length. The total length is divided among the total number of microtubules with some mean length such that Density =

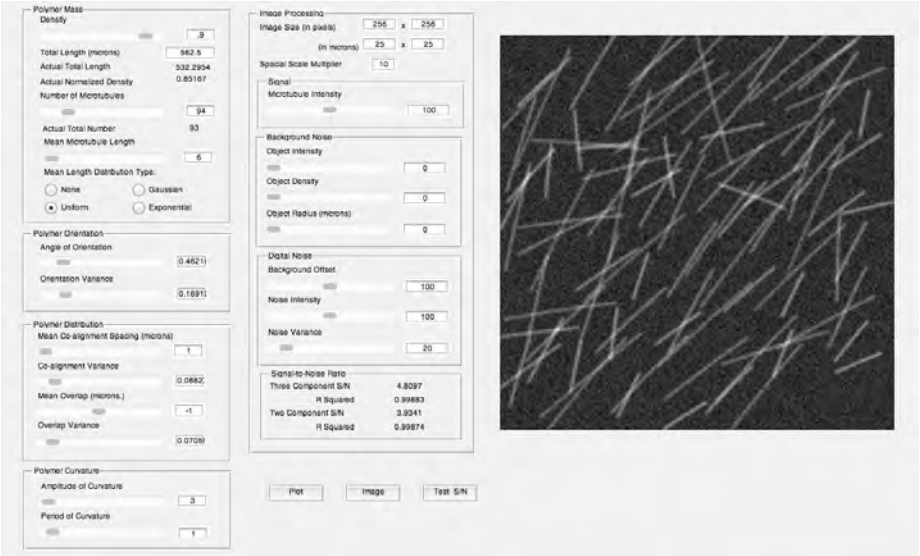


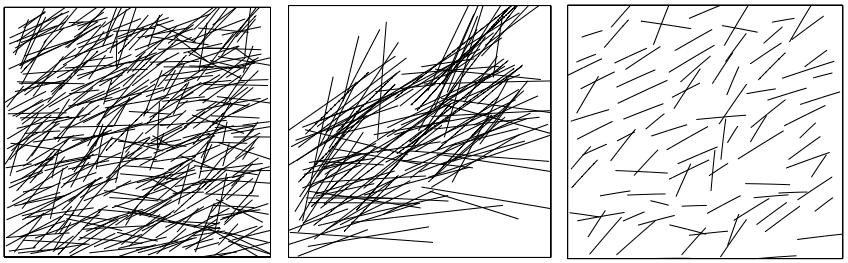
Figure 1: Screenshot of the simulator user interface showing an example configuration and resulting image.

Mean Length * Total Number. The resulting hyperbolic relationship between mean length and total number is incorporated into the simulator to preserve density such that adjusting one parameter modifies the other. Various density and length parameter combinations are shown in Figure 2. There is also the option to choose different types of distributions to vary microtubule length away from the mean including uniform, Gaussian, exponential, and no additional distribution.

Microtubules are not perfectly straight and generally have a slight bend or distortion. To add the ability to change the curvature a periodic function with variable amplitude, amplitude variance, period, and period variance, is added to a straight microtubule and the microtubule is truncated appropriately to preserve its allocated length. Figure 3 compares various microtubule curvature.

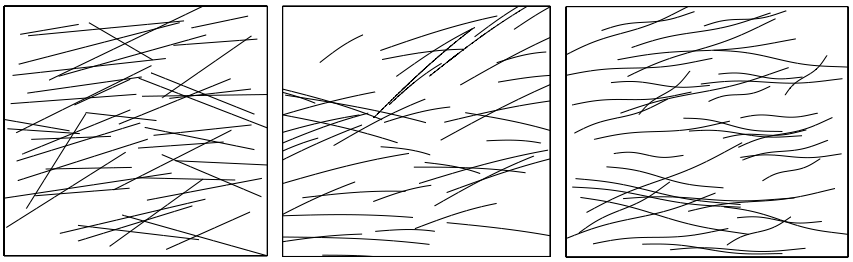
To describe the rotational or nematic order, the angle of orientation is a variable parameter in the interface with the ability to rotate the dominate angle by $\alpha \in [0, \pi)$. There is an additional parameter for angular variance that adds a Gaussian distribution with the given variance.

Two variables are required to describe translational or smectic order, one for each dimension. The alignment spacing t or the mean spacing along the director is controlled by the mean overlap parameter and an additional variance can be added with a Gaussian distribution. When $t > 0$, overlap is allowed between microtubule endpoints. When $t < 0$, the overlap parameter represents the required space between microtubule endpoints. The co-alignment spacing orthogonal to the director is described in a similar manner but must be non-



(a) (b) (c)

Figure 2: Various combinations of density and mean length parameters in 25μm by 25μm images. (a) and (b) have the same density but (a) has a shorter mean microtubule length than (b). (a) and (c) have the same mean microtubule length but (a) has a significantly higher density.



(a) (b) (c)

Figure 3: Various combinations of curvature amplitude and period in 25μm by 25μm images. (a) No curvature is added to the microtubules. (b) and (c) have the same low amplitude but (c) has a smaller period than (b).

negative. Several microtubule arrays with various order qualities are shown in Figure 4.

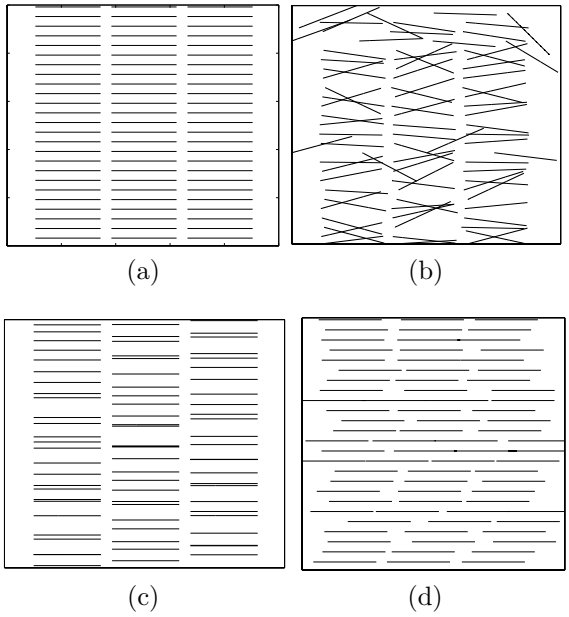


Figure 4: Several microtubule arrangements in 25μm by 25μm images created by varying order parameters. (a) is a perfectly ordered organization with no variance in any input order parameter. (b) varies the dominant angle of orientation but retains smectic order vertically. (c) varies the co-alignment spacing, s, but retains nematic order and smectic order into vertical columns. (d) varies the overlap spacing, t, but retains nematic and horizontal smectic order.

The microtubules are defined using parametric equations, one per microtubule, and modifies using the previously specified parameters. To create the image, the x and y coordinates of the microtubule equations are mapped to the indices i, j of a zero matrix M, respectively. The entry $M_{i,j} = M_{i,j} + 1$ such that intersections of microtubules are cumulative.

3.2.2 Shot Noise

In order to use florescence microscopy to image microtubules, Green Fluorescent Protein (GFP) is introduced into the system and binds to the microtubules. Viewed under specific wavelengths of light, GFP fluoresces and emits photons. By the nature of photon emission, the photons are emitted at random intervals which can be described by a Poisson distribution [6].

To account for this property of microtubule intensity, after the molecular distribution has been transcribed into an image matrix, the microtubule inten-

sities are assigned to have a mean intensity defined by the user interface with a Poisson distribution. The Poisson distribution can be approximated by a Gaussian distributed random variable with μ = mean intensity and $\sigma = \sqrt{\mu}$. The resulting microtubule intensity distribution can be seen in Figure 5. Looking at a cross section of the image in Figure 5, the microtubule intensity shown on the vertical axis varies significantly around the mean of 150 photons per pixel.

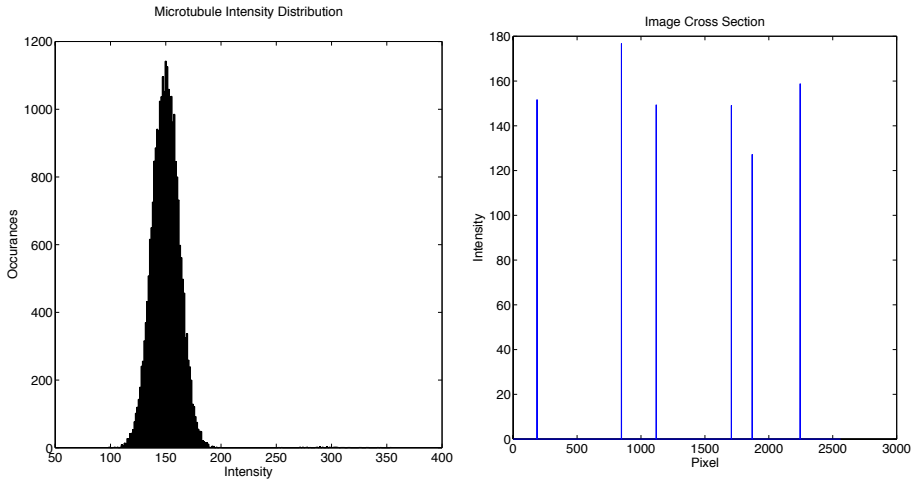


Figure 5: (left) The histogram of the image of microtubules after the addition of shot noise showing the Poisson distribution of microtubule intensities. (right) A cross section of the same image showing the wide distribution of microtubule intensities.

3.2.3 Diffraction

The aperture in a camera is the component of the lens that controls the amount of photons that enter the camera. When photons pass through the aperture, the photons are diffracted resulting in an airy disk. An airy disk is a series of concentric rings of different widths caused by diffraction through a 2-dimensional, circular aperture. The radius of the rings is proportional to the wavelength of light, λ, and inversely proportional to the numerical aperture of the lens, NA, such that the first ring, which contains 86% of the total photons, has a radius $r_{airy} = 1.22\frac{\lambda}{NA}$ [3].

To approximate the diffraction of photons into the airy disk, define a Gaussian filter with radius = r_{airy} where 5 standard deviations of the Gaussian filter fit in the disk. Using matrix convolution, apply the filter to the microtubule array image. The microtubule intensity distribution in Figure 7, has now been flattened by the Gaussian and no longer appears Poisson. A close up view of the microtubules at this stage in the process would show curves that are slightly blurred.

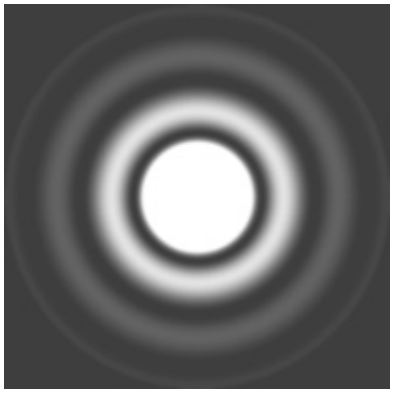


Figure 6: The airy disk pattern formed by the diffraction of photons through a 2-dimensional, circular aperture [2].

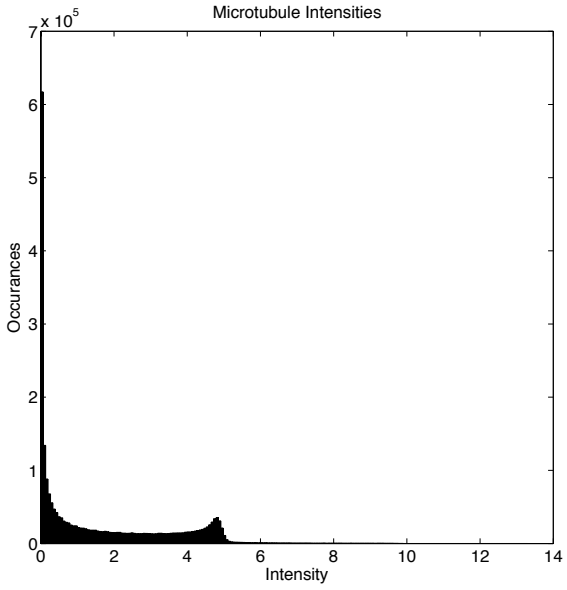


Figure 7: Histogram of microtubule array image showing the distribution of microtubule intensity after accounting for photon diffraction.

3.2.4 Digital Sampling

Cameras are restricted in resolution to the number of bin, or energy reservoirs, it contains. Photons that pass into the camera are diffracted and then captured in these bins. The amount of photons in the bin is recorded as intensity for an individual pixel. It may not be desired to have the final image be at the maximum resolution of the camera due to many considerations. To lower the resolution, the high-resolution image can be digitally sampled such that blocks of pixels are summed to create a smaller image matrix. The digitally sampled image D is given by

$$D_{i,j} = \sum_{j=1}^{n} \sum_{i=1}^{m} \sum_{k=1}^{s} M_{k+s*i,k+s*j}$$

where the resulting image I is $m \times n$ and the previous image M is $sm \times sm$ such that s is the spatial scale multiplier and the size of block used in the digital sampling.

This process only increases the range of intensities under the distribution but does not alter the microtubule intensity distribution in any other way as seen in Figure 8 since the shape of the distribution is preserved. For a more detailed analysis of this observation see Section 5.3 Challenge in Measuring Signal.

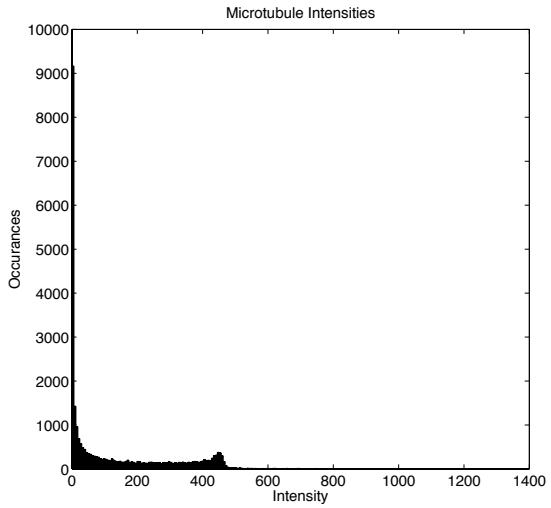


Figure 8: Microtubule intensities after digitally sampling the microtubule array image.

3.2.5 Camera Noise

To process the photons stored in the bins, the camera requires the use of electricity which influences the intensity readings adding a noise component. This noise, often referred to as digital noise, is a low-frequency noise and can be approximated with an appropriate offset and a Gaussian distribution defined by a noise variance parameter in the user interface. The histogram of the noise distribution is shown in Figure 9.

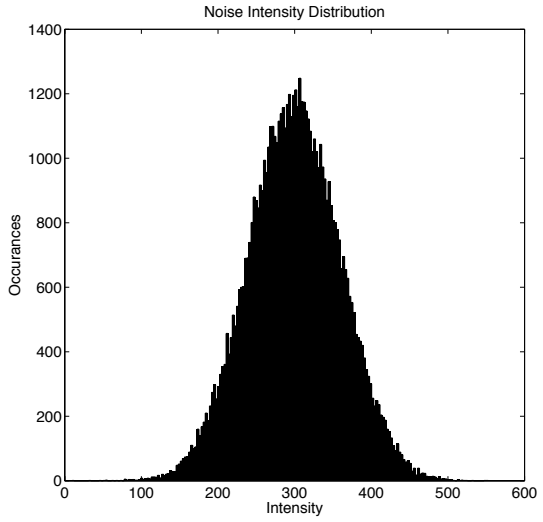


Figure 9: Distribution of the camera noise intensity with a variance of 60 photons and offset of 300.

3.2.6 Low-frequency Background Noise

An additional component is observed in live-cell images of microtubules. The concentration of GFP does not all bind with the microtubules; there is a proportion of the concentration that remains unbound and floating in the cytoplasm. This causes florescence not associated with any microtubule that appears as shapes in the background brighter than the general background noise.

To approximate this effect, an additional noise layer is added to the image where a set of circles are defined by parameters (density or number, radius, and intensity in the interface), dispersed throughout the image, and filtered through a similar process as the microtubule signal. One such possible combination is shown in Figure 10.

This situation is specific to florescence microscopy and poses challenges when analyzing microtubule images. Working with this noise component is difficult because of the resemblance to signal. Figure 11 compares the cross section of

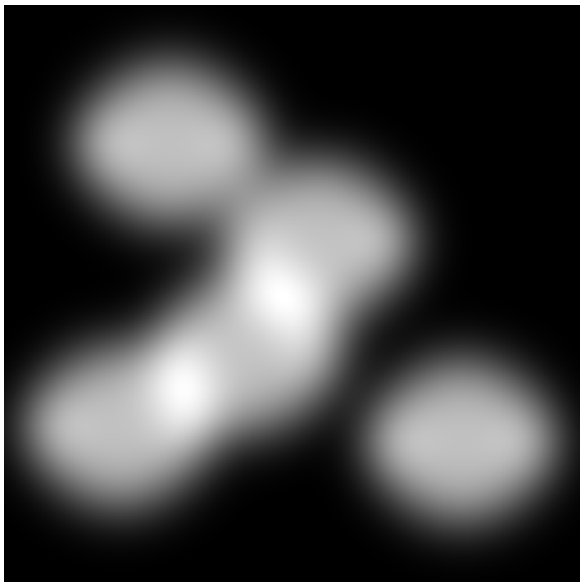


Figure 10: Image of the approximation of background noise from the florescence of unbound GFP that will be added to the final microtubule array image as noise.

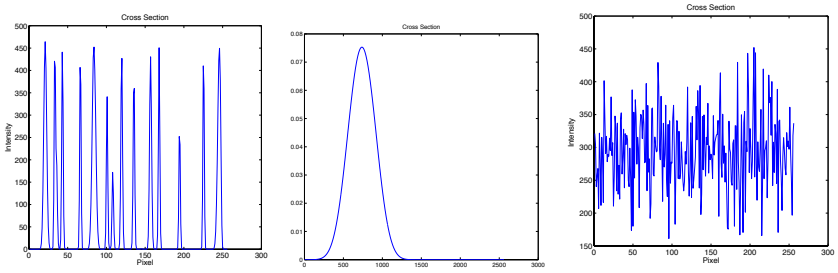
microtubule signal, background noise, and digital noise images. Notice that the cross section of the background noise is a low-frequency noise and resembles signal more than the traditional high-frequency noise. This makes it more challenging to identify.

3.2.7 Image Composition

To create the final image, the digitally sampled image, which is the final modification made to the microtubule intensity or signal, is summed with the noise components. The resulting image and respective histogram are shown in Figure 12.

4 Signal to Noise Ratio

An accurate measure of the signal to noise is needed in order to determine if the results of the algorithm are sufficiently confident to provide a measure of order. Given a low quality image there may not be enough separation between the measured microtubules and the rest of the image to make any sort of confident claim that the order measurements in fact measured microtubules and not phantom objects due to noise.



(a) (b) (c)

Figure 11: (left) Cross sections of (a) microtubule intensity or signal, (b) background noise, and (c) digital noise. (b) is a low-frequency noise and resembles the low-frequency signal (a) more than high-frequency noise (c).

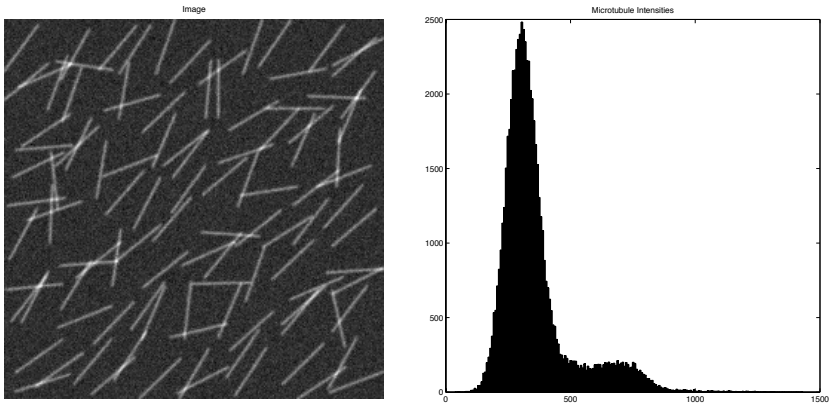


Figure 12: The final image (left) and the respective histogram (right) of the microtubule array with signal mean of 125, digital noise variance of 60, and no background noise.

4.1 The Signal to Noise Ratio as a Statical Measure

4.1.1 Z-Score

Expressing an observation as a z-score is a method of converting a distribution to a standard normal distribution such that the mean is 0 and the area under the distribution is 1. As a result a z-score is a statistical metric that is a standardized measure of the distance between an observation of a given distribution and the mean of that distribution where the units are in standard deviations of the standard normal distribution.

Definition 4.1 Given a population distribution with known mean and standard deviation, the z-score z is defined as

$$z = \frac{x - \mu}{\sigma}$$

where μ, σ are the mean and standard deviation of the population, respectively.

Definition 4.2 When only sample distribution statistics are known,

$$z = \frac{M - \mu}{\sigma_M}$$

where $\sigma_M = \frac{\sigma}{\sqrt{n}}$, the standard deviation of the mean, and σ and n are the sample standard deviation and sample size, respectively.

In this way, the z-score is a measure of the number of standard deviations the observed value falls from the mean.

Often the z-score is converted into a probability p of the observed value occurring. Since the distribution has been converted into the standard normal distribution which is a probability distribution function, the cumulative area under the curve represents the total probability that the observed value belongs to that region. A z-score can be written as the equivalent probability value p which is the area under the standard normal distribution between the mean and the observed value. For ease of use, it is common to look up the associated z-score and p values in a table.

4.1.2 Signal to Noise Ratio as a Z-Score

When analyzing images it is necessary to have an idea of how strong the signal is relative to the noise. Intuitively, the farther apart the signal and noise intensities are the easier it should be to distinguish signal from noise. Given a distribution of the intensities in an image with a known distribution of signal and noise, we can calculate the signal to noise ratio (S/N).

Definition 4.3 The signal to noise ratio (S/N) is defined as

$$S/N = \frac{\mu_{signal} - \mu_{noise}}{\sigma_{noise}} \tag{1}$$

where μ_{signal} = mean signal intensity, μ_{noise} = mean noise intensity, and σ_{noise} = standard deviation of noise intensities.

This is essentially a z-score calculation where μ_{sigma} is the observed value and the noise distribution is the population. Then S/N is the distance, or number of standard deviations, the signal mean is from the noise mean. The larger S/N is, the larger the distinction between signal and noise in the image.

4.1.3 Interpreting the Signal to Noise Ratio with a Z-Test

When looking at a region in an image that appears to be signal, how confident can we be that the observed pixel is in fact signal and not noise? What is the probability that the pixel is not noise? Is that probability small enough to accept potential for error? To determine a measure of confidence for the ability to distinguish signal from noise in the image apply a statistical z-test.

Certain conclusions can be drawn from an observation with a given z-score using a z-test, which is comparable to the students t-test except μ and σ of the distribution are known and not estimated. The z-test can only be used when the population distribution is known.

Given an image with known noise and signal distributions, the S/N value of the image is used as the z-score in the z-test. The hypothesis being tested is that the measured intensity is noise and not signal. Hence define H_0 as the measured mean signal intensity = the mean noise intensity. For the significance level, let $\alpha = .01$ to ensure 99% confidence in the determination. This yields critical z-score values at $z = \pm 2.58\%$ at which the associated $p = .01$. If $S/N > 2.58$, the null hypothesis is rejected and we can assume with 99% confidence that the measured intensity is not noise. If $S/N < 2.58$, then the null hypothesis is accepted and we cannot be confident enough to conclude that the measured intensity is not noise.

4.1.4 Confidence Level and Margin of Randomness

In general when taking the S/N as a z-score and determining the associated p value, the confidence of the measured intensity not being noise is given by $(1-p)*100$. This forces images where signal can be distinguished from noise with 95% and 99% accuracy to have a minimum S/N of 1.96 and 2.58 respectively.

Given an image that is 512 by 512 pixels in size with a 99% confidence in signal detection, then the amount of uncertainty or random events still predicted in the image is 1% of the number of pixels in the image which in this case is 2621 pixels. This implies that the larger the resolution of the image the more error is allowed at the same confidence level and a higher confidence level and S/N may be desired.

5 2-Component Mixed Gaussian as Method for Calculating the Signal to Noise Ratio

Using the MATLAB simulator that constructs a representative image by adding the individual signal and noise layers together, the exact S/N of the simulated

image can be calculated. Because of this ability to access a pure signal and pure noise image, the accuracy of different techniques in calculating the S/N can be compared to the true S/N.

5.1 Determining the Predicted Signal-to-Noise Ratio

The final image displayed by the simulator is created by summing an image of pure signal and pure noise. To calculate the true or predicted S/N, determine the signal mean, noise mean, and noise variance and then apply Equation 1.

To calculate the predicted signal mean, compute the mean intensity of the pure signal distribution given by the nonzero pixel intensities of the histogram of the microtubule array image after the digital sampling. Computing the average and standard deviation of this distribution yields the best fitting Gaussian to the signal image histogram computed by other conventional measures.

For images with only one noise source from the digital imaging process, use the non-zero pixels of the true noise intermediary image and compute the mean and standard deviation of the noise distribution. For images with two noise sources, calculate the mean and standard deviation of both noise components independently. The combined noise mean and standard deviation are given by

$$\mu_{noise} = \frac{\mu_1 + \mu_2}{2}$$

$$\sigma_{noise} = \sqrt{\sigma_1^2 + \sigma_2^2}$$

where μ_1, σ_1 and μ_2, σ_2 are the mean and standard deviation of the digital and low-frequency background noise components, respectively.

Applying Equation 1, the true S/N can be calculated and used as the predicted value in testing.

5.2 Determining Signal and Noise in Images

To be precise the S/N should be calculated from two images: one that is pure signal and the other signal with noise added. This is not a possible situation with florescence microscopy since the noise is from the imaging process itself and pure signal cannot be captured. In this field, a widely used method is to select a region in the image that represents purely background and thus the noise and a separate region from the same image that is a good estimation of signal and noise. We would like to instead determine the accuracy of an autonomous method.

Given just a single image with both signal and noise components, the histogram can be used to approximate the signal and noise distributions and segment the image into two parts. Through the process of simulating images of microtubule arrays, it is evident that the pure signal component and pure noise component can be described approximately with Gaussian distributions. The

signal and noise components are added together which implies that the final image can be described using a weighted sum of two Gaussian distributions given by

$$g(x) = p * \frac{1}{\sigma_n\sqrt{2\pi}} e^{-\frac{(x-\mu_n)^2}{2\sigma_n^2}} + (1-p) * \frac{1}{\sigma_s\sqrt{2\pi}} e^{-\frac{(x-\mu_s)^2}{2\sigma_s^2}} \quad (2)$$

where p is the proportion of the first Gaussian in the total mixture, σ_n, μ_n are the standard deviation and mean of the noise component and σ_s, μ_s are the standard deviation and mean of the signal component.

Using MATLAB to fit Equation 2 to the histogram of image intensities using the Maximum Likelihood Estimation technique, the two Gaussians can be fit to the histogram where the first Gaussian approximates the pure noise distribution and the seconds approximates the pure signal distribution.

5.3 Challenge in Measuring Signal

As seen in Figures 5, 7, and 8 the width of the signal distribution fluctuates during the imaging process. Given parameter values such that the spatial scale is 10 pixels for a 256 × 256 pixel image representing 25μm square and an input microtubule intensity is 100 photons, after the filter simulating diffraction has been applied the measured microtubule intensity or signal is reduced to 7.6293 photons which is only 7.5% of the initial input. After the image is created and is digitally sampled, the measured signal has resumed more of the original value and the predicted signal is determined to be 29.6099, only 29% of the original value.

This compromising result is influenced by the spatial scale used which has an effect on both the Gaussian filter size and the size blocks used in the digital sampling.

Size	Filter Size	Input	After Diffraction	After Digital Sampling
2	9.0000	101.6445	7.6293	29.6099
4	19.0000	101.5297	3.6140	50.5958
6	28.0000	101.0212	2.3627	73.5824
8	37.0000	100.8082	1.8013	98.6161
10	46.0000	100.8611	1.4250	122.2494
12	56.0000	100.9799	1.1737	144.5578
14	65.0000	100.9743	1.0248	171.1403

Table 1: Signal compromise through the processes of diffraction and digital sampling.

The effects of the spatial scale can be seen on the signal intensity in Table 1 and on the percent of original signal input in Table 2. The larger the spatial scale the more drastic the diffraction filter compromises the signal and larger

Size	Filter Size	Input	% After Diffraction	% After Digital Sampling
2	9.0000	101.6445	7.51	29.13
4	19.0000	101.5297	3.56	49.83
6	28.0000	101.0212	2.34	72.84
8	37.0000	100.8082	1.79	97.83
10	46.0000	100.8611	1.41	121.21
12	56.0000	100.9799	1.16	143.16
14	65.0000	100.9743	1.01	169.49

Table 2: Percent signal compromise through the processes of diffraction and digital sampling

recovery by the digital sampling. Since the spatial scale is proportional to the size of the filter, the increase scale increases the filter size and thus the area over which the pixel intensity is dispersed. Similarly, the larger spatial scale increases the pixel area which is summed during the digital sampling process.

The resulting effect is predicted signal intensity that is proportional to the input microtubule intensity but is an underestimation of the true signal input as shown in Figure 13. The line fit to the correlation of input and measured signal is .815 of the input intensity. Accounting for this compromise in signal when creating the image is challenging and is outside the scope of this project. The signal in the image after digital sampling is the signal used to generate the final image so the measured intensity of that image is assumed to be the 'true' signal of the image. To ensure a specific signal intensity in an image, determine the factor by which the signal is compromised and account for that factor in the microtubule intensity parameter of the user interface.

5.4 Calculation Error

To test the error in calculating the S/N, limit the parameter space to two dimensions. Let the signal or microtubule intensity range from 20 to 170 by intervals of 30 and let the noise variance range from 10 to 90 by steps of 20. After measuring 10 samples at the 24 different signal-noise parameter combinations, the percent error is shown in Figure 14.

The calculation of S/N using a 2-Component Mixed Gaussian, Equation 2, has a large systematic error shown by all errors residing in a band between 51% and 84% error. In addition to the systematic error, there is also an additional spread within the 33% error band. Figure 14 shows three distinct regions with separate trends: $S/N \leq 1.5, 1.5 < S/N \leq 5$, and $S/N > 5$.

When the $S/N \leq 1.5$, the calculation is inconsistent spanning the full height of the error band. Interpreting a S/N = 1, there is only a 68% confidence that the observed intensity is signal and it is not surprising that the calculation results in large discrepancies when trying to measure signal given that it is

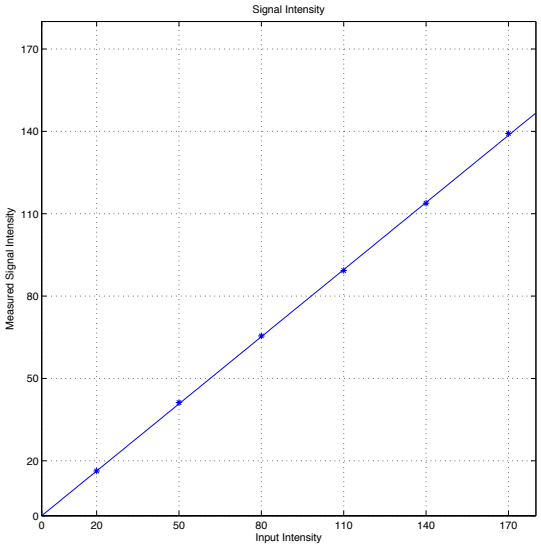


Figure 13: Correlation between the input and measured signal intensity of the image where the input intensity is specified in the user interface and the measured intensity is the mean of the microtubule intensities of the digitally sampled image. The measured intensity is .815 of the input intensity.

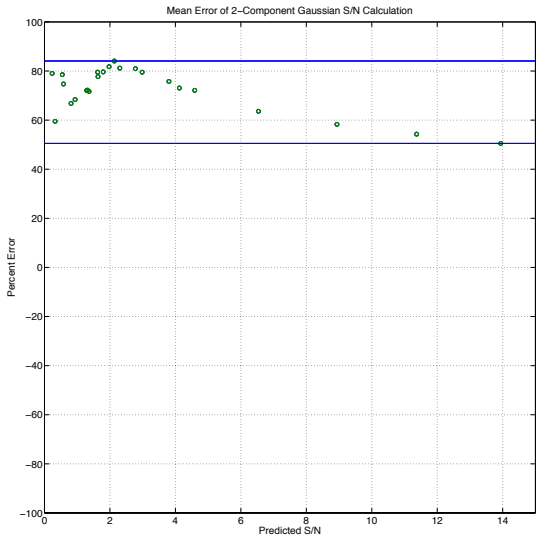


Figure 14: Mean error of the 2-Component Mixed Gaussian S/N Caluclation with error bands at 51% and 84%.

difficult to distinguish from noise at low S/N. This property is made clear in the histogram of intensities in the image with a S/N = 1.30 shown in Figure 15 with microtubule intensity = 100 and noise variance = 100. It is very unclear where the second Gaussian is since the noise or first Gaussian dominates the image.

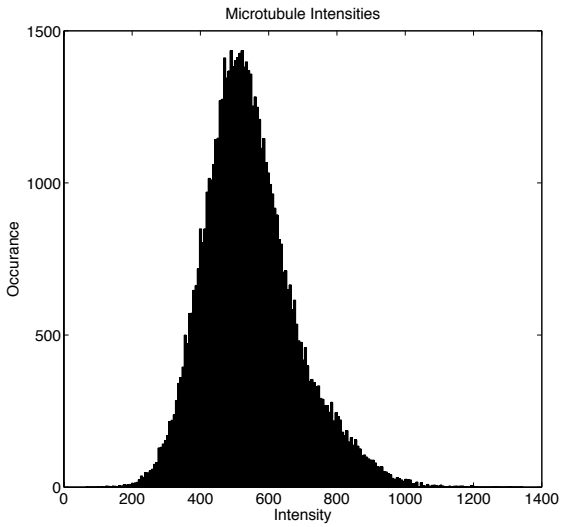


Figure 15: Histogram of final image with S/N = 1.30, microtubule intensity = 100, and digital noise variance = 100. The signal intensity is no longer evident in the histogram.

When $1.5 < S/N \leq 5$, this is the optimal range for the algorithm because of the clustering of error within the band. After adjusting for the systematic error, this range of S/N would consistently provide the most accurate measurements. This range also comprises the realistic observed S/N noise in images and accuracy is desired most in this range.

S/N ratios above 5 are uncommon and ratios above 10 are rare for normal images that can be captured with light microscopy. The percent error is linearly descending and would seem to be improving the accuracy. But when accounting for the systematic error in the calculation, the decreasing percent error less than 0 is generating more error.

5.4.1 Signal Error as the Source of the Systematic Error

The largest contributor of error in the calculation of S/N is the error in calculating the mean signal intensity. The mean signal error is directly correlated with the total calculation error. Visibly the mean signal error shown in Figure 16 follows the same general trend as the S/N error in Figure 14.

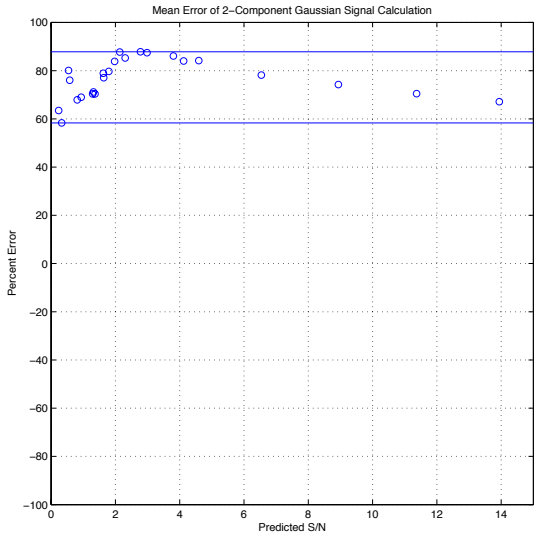


Figure 16: Mean error of the 2-Component Mixed Gaussian calculation of signal and maximum and minimum error shown by the error bands at 58% and 87%.

Generated from the same data set as Figure 14, there is a slightly larger systematic error and spread, for the percent error resides in a band between 58% and 87%. This suggests that the largest source of error is from the mean signal calculation and only a small portion of the mean signal error is compensated for by the remaining calculation.

The predicted mean signal intensity is highly correlated to the observed mean signal intensity, as shown in Figure 17, and the observed signal intensity is roughly proportional to the predicted signal intensity (approximately 1.8 times larger shown by the trend line in Figure 17). An error trend that is proportional to actual values is equivalent to a systematic error in relative error. As such, this proportional error in the measurement of signal mean is the cause of the large systematic error dominating the S/N calculation.

The error in measuring mean signal intensity is a specific fault of the 2-Component Gaussian fitting technique. When adding the low-frequency digital noise, the addition of the offset aligns the lowest signal intensities with the mean of the noise component. But since the signal and noise are additive, the signal intensities that occur at the same intensities as the noise distribution are consumed by the noise distribution adding a slight skew to the noise Gaussian distribution that is not visible in a histogram given the magnitude of noise. These low intensity signal values are then calculated as noise and not as signal which forces the measured signal mean to be larger than the true mean. Figure 18 shows a plot of the histogram of an image with both the 2-Component Mixed Gaussian fit determining the measured mean and the Gaussian fit to the pre-

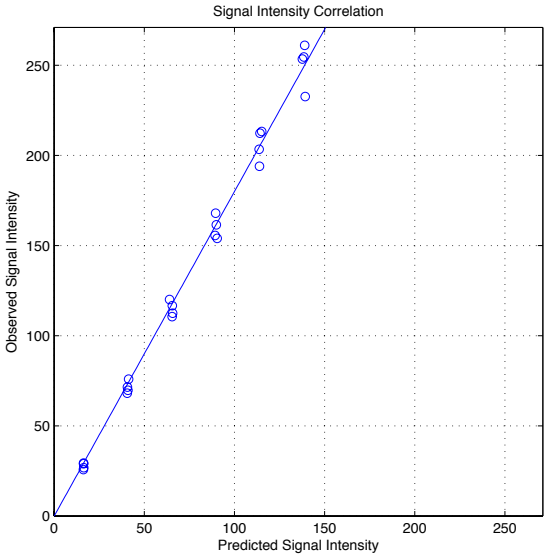


Figure 17: Correlation between the predicted and observed signal intensity showing a 1.8 overestimation in the calculation of the signal mean.

dicted signal. Note the predicted Gaussian fit is justified below the measured Gaussian representing signal by the technique such that the predicted signal intensity is 154.3 and measured signal intensity is 345.9 after accounting for offset with noise variance at 60 resulting in a systematic error.

A 60% systematic error is not an acceptable amount of error for the purpose of the S/N calculation. Assuming a desired 99% confidence in distinguishing signal from noise and the actual S/N = 2, the image should not be used since a S/N = 2 as a z-score only guarantees 95% confidence. But, the 2-Component Mixed Gaussian fit calculation with a 60% systematic error used on the same image measures the S/N = 3.2 which corresponds to a confidence score of 99.9%. As a result a Type II error occurs and the image is falsely considered to satisfy the desired confidence level.

5.5 2-Component Gaussian Error of 2-Source Images

Florescense microscopy images of microtubule arrays have the additional low-frequency noise which distorts the right half of the noise Gaussian shown in Figure 19.

Using the same method of test from the 2-Component Mixed Gaussian on 1-source images, the technique was analyzed using 2-source images where the mean background noise intensity was 50 in the first run and 100 in the second, the mean error of calculating the S/N was determined. The resulting percent error is shown in Figure 20.

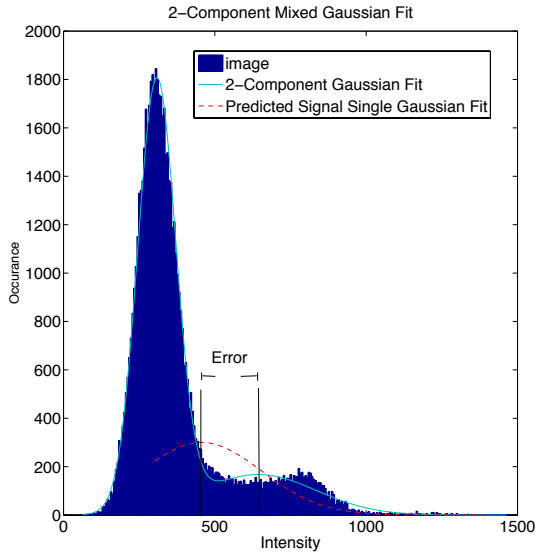


Figure 18: Histogram of microtubule array image with 2-Component Mixed Gaussian Fit in light blue and actual signal Gaussian fit in dashed red. The difference between the peaks is the error in signal mean measurement.

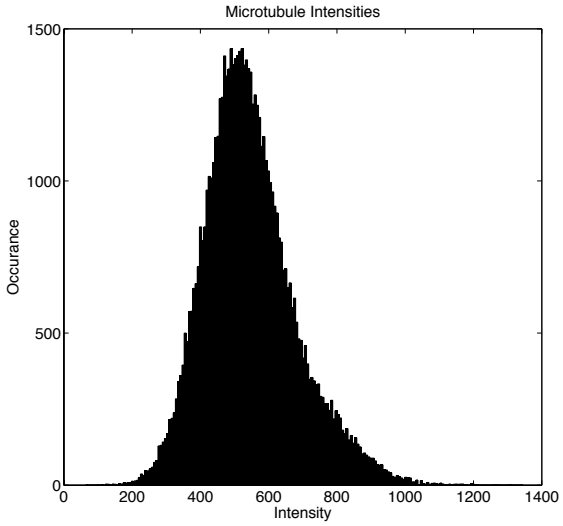


Figure 19: Histogram of a microtubule array image that has two sources of noise, both digital noise and low-frequency background noise, resulting in a slight skew in the right of the large noise distribution.

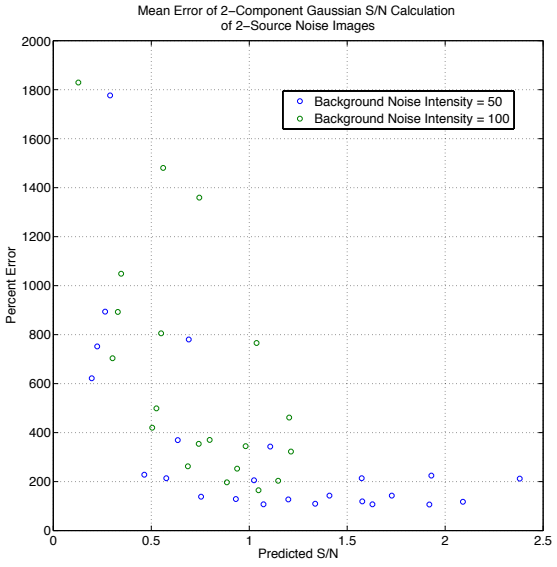


Figure 20: The percent error of the S/N calculation using the 2-Component Mixed Gaussian technique on images with 2 sources of noise. Two trials are shown at different intensities of the additional background noise source.

Notice that with the additional noise source even at a low level of intensity the value of the S/N is significantly reduced. Looking at the region of interest of valid S/N between [1.5, 2.5], the error of calculating the S/N has approximately doubled. The same calculation error in determining signal error holds and is only made worse by the additional noise component.

As a result of this simple analysis, it is clear that calculating the S/N of microtubule images is more difficult than with other images. The additional noise factor adds another layer of complexity and compounds the error inherent to the 2-Component Mixed Gaussian technique. A more complex method for determining the S/N autonomously is required for images of microtubule cytoskeletal arrays.

5.6 Technique Modifications for Improved Accuracy

There are several modifications to the 2-Component Mixed Gaussian technique that may show improved accuracy. With the addition of a second source of noise which also has a Gaussian distribution of intensities, a mixed Gaussian with three components could be fit to the histogram to try and fit a Gaussian distribution to both sources of noise individually.

More ad hoc methods could be devised such as fitting the noise distribution to only the right side of the noise distribution in the histogram which becomes

skewed with the addition of low-frequency background noise. Additionally, since the source of error is often the determination of signal mean, the signal mean could be determined using another technique such as analyzing a cross section of the image and only try and fit Gaussian distributions to the noise components.

6 Considerations for Methods of Determining Order Parameters

When considering possible techniques for determining order parameters in microtubule array images, the affect noise in the image has on the calculation should be a priority. For example, the Radon Transform which is a promising technique for determining order parameters, is relatively not sensitive to low-frequency noise. The Radon Transform is fundamentally a line integral which causes the high-frequency noise to cancel with itself [7]. Additionally, steps should be taken to consider the effects of low-frequency noise on the Radon Transform and any other potential algorithms.

After a method has been defined, testing using images of known S/N will be able to determine the range of S/N an input image must have in order to produce a confident measure of order.

References

[1] ALCOM at Kent State University & Case Western Reserve. (2004). *Polymers & liquid crystals.* Retrieved from http://plc.cwru.edu/tutorial/enhanced/files/textbook.htm

[2] Cambridge in Colour. (2013). *Lens diffraction & photography.* Retrieved from http://www.cambridgeincolour.com/tutorials/diffraction-photography.htm

[3] Davidson, M. W. (2013). *Numerical aperture and image resolution.* Retrieved from http://www.microscopyu.com/tutorials/java/imageformation/airyna/

[4] Deinum, Eva E.. (2013). *Simple models for complex questions on plant development.* Ph.D. Thesis. Wageningen University: Netherlands.

[5] Liquid Crystal Institute at Kent State University. (1999). *About liquid crystals.* Retrieved from http://www.lci.kent.edu/lc.html

[6] Optical Technologies. (2008). *Noise in photoreceptiors.* Retrieved from http://optical-technologies.info/tag/shot-noise/

[7] Pourreza, R., Banaee, T., Pourreza, H., & Kakhki, R. D. (2008). *A radon transform based approach for extraction of blood vessels in conjunctival images.* Retrieved from http://profdoc.um.ac.ir/articles/a/1007402.pdf

Prime Factorization of Kászonyi Numbers

Ariana Cappon and Emily Walther

Abstract

Snarks are a class of simple, cubic, non-planar graphs that cannot be edge-3-colored. By a result of Kászonyi, if G is a snark, e is an edge of G, and G_e is the cubic graph that one obtains by deleting the edge e and "eliminating" its endpoint vertices, then the number of edge-3-colorings of G_e with three given colors will be $18 \cdot \psi(G, e)$ for some nonnegative integer $\psi(G, e)$. It has been previously shown that there exists a cyclically 4-edge connected snark G_0 with and edge g_0 such that $\psi(G_0, g_0) = 2^a \cdot 3^b \cdot 5^c \cdot 7^d$ where a, b, c, and d are arbitrary non-negative integers. In this note, we will show that for every positive integer n where prime factors of n are all less than or equal to 149, there exists a snark G and an edge e of G such that $\psi(G, e) = n$.

1 Introduction

In 1852, Francis Guthrie was coloring a map of England when he noticed that he only needed four colors in order for two bordering countries to not be the same color. This observation intrigued mathematicians and led to the Four Color Problem and eventually the Four Color Theorem, which was proved in 1976. The four color theorem states that given any separation of a plane into finitely many contiguous regions, no more than four colors are required to color the regions of the map so that no two adjacent regions have the same color. For a more detailed history and explanation of the theorem, see the book by Wilson [Wi]. The four color problem (later the Four Color Theorem) led mathematicians to study further colorings of graphs, including "edge-3-colorings". A question soon arose: do cubic, bridgeless, non-planar graphs exist that cannot be "edge-3-colored"? The answer was soon found to be yes, and the term "snark" was born (borrowed in [Ga] from *The Hunting of the Snark* by Lewis Carroll) to name this new class of graphs. In this paper, we primarily will be exploring the prime factorization of the Kászonyi numbers of snarks and their edges.

1.1 Some Preliminary Graph Definitions

In this section we will give a brief survey of the definitions and terminology relevant to our research. For our purposes, the graphs we refer to will be undirected, with no loops or multiple edges.

Definition 1.1 Consider a graph G that has finitely many vertices, no loops, and no multiple edges. We will denote the set of edges in G as $E(G)$, and the set of vertices in G as $V(G)$.

Definition 1.2 In a graph G, a vertex is "n-valent" if it is connected to exactly n edges. A vertex connected to exactly one edge is called "univalent."

Definition 1.3 Given a graph G that is finite with no loops or multiple edges:
1. G is "cubic" if all of its vertices are 3-valent.
2. G is "quasi-cubic" if all of its vertices are either 3-valent or univalent.

Definition 1.4 Suppose G has finitely many vertices, no loops and no multiple edges. Say there exists a subgraph of G with distinct vertices $v_1, v_2, \ldots, v_n$, and edges $(v_i, v_i + 1)$, $i \in \{1, 2, \ldots, n-1\}$ and (v_n, v_1), for n$\geq$1. This subgraph is called a "cycle" in G. An "n-cycle" has exactly n vertices.

Definition 1.5 Suppose G is a finite graph with no loops or multiple edges. The "girth" of G is the number of vertices in a cycle of G with the smallest number of vertices.

Definition 1.6 A graph is "simple" if is has finitely many vertices, no loops, no multiple edges, and is *connected*.

Definition 1.7 Two or more graphs are pairwise disjoint if they share no edges or vertices.

Definition 1.8 If G is a simple graph, and e is an edge of G, then $G-e$ denotes the graph where e is removed, while the vertices connected to e are retained. This process is called *edge elimination*.

Definition 1.9 If G is a simple graph, and v is a proper subset of $V(G)$, then $G - v$ denotes the graph where all elements of v are removed, and all edges connected to v are also removed. This process is called *vertex elimination*.

Definition 1.10 Consider the simple graph G with vertices u, v, u_1, u_2, v_1, v_2, and edges (u, v), (u, u_1), (u, u_2), (v, v_1), (v, v_2). (For convenience, we will rename edge (u, v) as e). To *subtract* e, eliminate e and eliminate vertices u and v. Then, create two new edges, (u_1, u_2) and (v_1, v_2). We call the resulting graph G_e. This process is called *edge subtraction*.

Definition 1.11 A simple graph G is at least "cyclically n-edge connected" if no two cycles of G can be separated from each other by the elimination of at most $n - 1$ edges not in either cycle.

1.2 Edge-3-Coloring of Graphs

Definition 1.12 Suppose G is a simple, cubic graph. Let a, b, and c be the non-zero elements of $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ and let $E(G)$ be the collection of all of the edges of G. G is colorable (edge-3-colorable) if there exists a function: $\gamma : E(G) \to \{a, b, c\}$, such that for any two adjacent adges e_1 and e_2 of G, $\gamma(e_1) \neq \gamma(e_2)$.

Definition 1.13 Suppose G is a simple, cubic graph. Let $E(G)$ be the collection of all of the edges of G. An "edge-3-decomposition" of G is a partition of $E(G)$ into 3 different classes such that if e_1 and e_2 are adjacent edges, they are not in the same class.

Note that the cardinality of the set of all edge-3-colorings will always be 6 times the number of edge-3-decompositions due to the 6 permutations of the colors a, b, and c.

Definition 1.14 Suppose G is a simple, cubic graph and γ is an edge-3-coloring of G. A "Kempe chain" (in fact a "Kempe cycle") for γ is a subgraph K of G which is maximal with respect to having the following two properties:
(i) K is connected.
(ii) The edges of K have just two colors (under γ).

Remark: If the simple graph G contains a Kempe chain K, assigned the colors x and y, then all edges connected to K but not in K must be assigned a third color, z. Thus, the coloring of the edges within K can be interchanged, so that all edges within K colored x are now colored y, and all originally colored y are now colored x. This color change does not affect the rest of G, because all edges connected to but not in K are still colored z.

Lemma 1.15 *(Parity Lemma) Let G be a simple, quasi-cubic, colorable graph containing at least one 3-valent vertex. Given any coloring of G, the number of univalent vertices of G connected to an edge colored x is even if there are an even number of univalent vertices in G. Similarly, the number of univalent vertices of G connected to an edge colored x is odd if there is an odd number of univalent vertices in G.*

The above lemma is equivalent to the following statement:

Lemma 1.16 *(Parity Lemma) Let G be a simple, quasi-cubic, colorable graph containing at least one 3-valent vertex, and the set of edges connected to univalent vertices $\{e_1, e_2, \ldots, e_n\}$. Let γ be an edge-3-coloring of H. Then*

$$\sum_{i=1}^{n} \gamma(e_i) = 0 \text{ (the zero element of } \mathbb{Z}_2 \oplus \mathbb{Z}_2).$$

From the second version of the parity lemma arises the idea that the edge-colorings of a minimal cut set must add to 0 in $\mathbb{Z}_2 \oplus \mathbb{Z}_2$.

In 1880, Peter Tait was one of the first mathematicians to explore the"edge-colorings" of graphs (for more details on Tait, see [Wi]). Tait soon noticed that there was a connection between the "face-colorings" of a simple, planar, cubic, bridgeless graph and the ways you could color the edges of that graph.

Theorem 1.17 *A simple, cubic, planar, bridge-less graph can be:*
i.) face-4-colored
ii.) edge-3-colored

Note that an "edge-3-coloring" of such a graph directly results from a "face-4-coloring" of a graph in a non-trivial way (for details, see [Wi]).

2 Defining Snarks

We will now introduce a family of graphs that are almost exclusively examined in the rest of this paper.

Definition 2.1 A snark is a simple, cubic, graph G such that:
1. G cannot be 3-edge-colored,
2. The girth of G is at least 5,
3. G is at least cyclically 4-edge-connected.

Remark on the conditions that a snark must satisfy:
i) The girth of G is at least 5.
A non-colorable cubic graph that contains a 3-cycle or a 4-cycle can be reduced in a trivial way to a smaller non-colorable cubic graph.
ii) G is at least cyclically 4-edge-connected.
If G is non-colorable and only 1-edge connected, then G contains a bridge and consequently G is non-colorable by an elementary argument. If G is non-colorable and only 2 or 3-edge-connected, then at least one of the minimal cut sets (resulting in disconnected, cubic graphs) must be non-colorable (after the cut ends are tied up in an appropriate, simple manner).

One famous example of a snark is the Petersen Graph, constructed by Julius Petersen in 1891. This graph is the smallest snark and is notable for its various symmetries. We will examine the Petersen graph further throughout this paper.

2.1 Ways to construct larger snarks from smaller snarks

There are two relevant methods for constructing arbitrarily large snarks. Here we will describe them briefly. The first is the dot product, a construction of Isaacs [Is], and the second is superposition, developed by Kochol [Ko]. Isaacs' dot product involves cutting and connecting the edges of two snarks to form a bigger snark. A superposition involves replacing an edge of a snark with another snark to form a bigger snark. (Note that the dot product can be represented as a special case of a superposition.) This construction was particularly useful for Scott McKinney's research, and is fundamental to the results of this paper.

2.2 McKinney's Coloring of a "Ripped" Petersen Graph

Definition 2.2 A star is a set of 5 vertices, denoted $v_1....v_5$, of a snark G such that there is always an edge between v_i and v_{i+2} and an edge between v_i and v_{i+3} where + is addition in $\mathbb{Z}_5$. Let u_i denote the vertex in G that is connected to v_i but is not v_{i+2} or v_{i+3}.[McK]

Theorem 2.3 *(Star Coloring Theorem) Let G be a snark and e be an edge of G. Let G_e be colorable. For some $i \in \mathbb{Z}_5$, the colors of the edges (v_i, u_i),...,(v_{i+4}, u_{i+4}) are x, y, x, x, and z (respectively) for some choice of distinct colors x, y, and z from $\{a, b, c\}$. [McK]*

Note that each Petersen graph contains a star. Following the Star Coloring Theorem, Scott McKinney gave an important lemma about coloring a specific "ripped" Petersen Graph. The specific "ripped" Peterson graph is as follows:

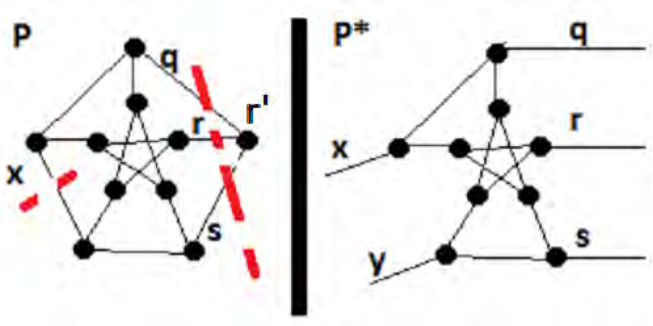


Figure 1: A "Ripped" Petersen Graph.

Definition 2.4 Let x and r be two edges of a Petersen graph P such that x and r share no adjacent edges (see Figure 1). Without loss of generality, let r' be a vertex of r. Let edges q and s be the other two edges connected to r'. Let P^* be the graph that results from "snipping" edge x (extending the two "loose ends" into two new edges x and y) and removing vertex r' (extending the "loose ends" q, r, and s).

Lemma 2.5 *(McKinney) Suppose that in the "Ripped" Petersen graph P^* in Figure 1, the edges x, y, q, r, and s are each assigned a color in $\{a, b, c\}$ such that*
(i) the five assigned colors add up to 0 (in $\mathbb{Z}_2 \oplus \mathbb{Z}_2$) and
(ii) the colors assigned to the edges x and y are distinct.
Then these five colors extend to a unique coloring of the entire "Ripped" Petersen graph P^*.

Note that in any possible coloring of P^*, x and y will never be assigned the same color:

Proof For the sake of contradiction, assume a coloring of P^* exists where x and y are assigned the same color. Then q, r, and s would all have to be assigned different colors (Parity Lemma, applied to edges q, r, s, x, and y). This would result directly in a coloring of the Petersen Graph which is a contradiction because the Petersen Graph is a snark. □

3 Theorems of Kászonyi

Our research stems from results of László Kászonyi, presented in this section. Definition 3.1 and Theorems 3.2, 3.3, and 3.5 below all came from the papers of Kászonyi [Ka1, Ka2, Ka3]. For a convenient exposition (including proofs), see [Br2, section 3].

Definition 3.1 Suppose H is a simple, cubic, edge-3-colorable graph with edges d_1 and d_2. These edges are "orthogonal" if there does not exist $\gamma \epsilon EC(G)$ for which d_1 and d_2 are edges in the same Kempe cycle.

Theorem 3.2 *(Kászonyi) Suppose G is a snark and e is an edge of G. Let d_1 and d_2 be the edges of G_e that result from removing edge e (i.e. $d_1 = (u_1, u_2)$ and $d_2 = (v_1, v_2)$ in Definition 1.10). If G_e is colorable, then d_1 and d_2 are orthogonal edges.*

Theorem 3.3 *(Kászonyi) If G is a snark and e is an edge of G, then card $EC(G_e) = 18L$ for some nonnegative integer L.*

Definition 3.4 For a given snark G and a given edge e, the Kaszonyi number of G and e is the integer L in Theorem 3.3 and is denoted $\psi(G, e)$.

Remark: The factor of 18 in Theorem 3.3 comes from two seperate, smaller factors. A factor of 6 comes from the 6 permutations of the colors a, b, and c; and the remaining factor of 3 is a direct result of the fact that the resulting edges d_1 and d_2 referred to in Theorem 3.2 are orthogonal, i.e. cannot be in the same two-color Kempe cycle.

Theorem 3.5 *(Kászonyi) Given the Petersen graph P and any edge e of P, $\psi(P, e) = 1$.*

4 Our Main Results

Definition 4.1 Let $\mathscr{S}$ be the following set of prime numbers, defined in three steps:

$\mathscr{S}_1 := \{p \epsilon \mathbb{N} |\ p \text{ is prime and } p \leq 149\}$
$\mathscr{S}_2 := \{173, 179, 181, 197, 229, 257, 271, 359\}$
$\mathscr{S} := \mathscr{S}_1 \cup \mathscr{S}_2$

Theorem 4.2 *Suppose n is in the form:*

$$n = \prod_{p \epsilon \mathscr{S}} p^{m(p)}$$

where for each $p \epsilon \mathscr{S}$, $m(p)$ is a non-negative integer. Then a snark G and an edge e of G exist such that $\psi(G, e) = n$.

This theorem will be based on the following lemma:

Lemma 4.3 *Suppose $p \epsilon \mathscr{S}$ and n is a positive integer such that for some snark G_0 and some edge e_0 of G_0, $\psi(G_0, e_0) = n$. Then there exists a snark G and an edge e of G such that $\psi(G, e) = p \cdot n$.*

Theorem 4.2 follows from Lemma 4.3, Theorem 3.5, and induction. For $p = 2, 3$, Lemma 4.3 was shown by applying the results of Dr. Richard Bradley (see [Br1] Theorem 2.2 and let the two "smaller" snarks there be G_0 and the Petersen Graph). The cases where $p = 5, 7$ were shown by Scott McKinney (see [McK] Theorem 5.1 and Theorem 5.3). Our task is to show the cases where $p \geq 11$.

To illustrate the main idea of the proof, we shall first give the argument twice, in different ways, for $p = 17$.

4.1 Background information for both proofs

In order to do this, we add additional edges onto Scott McKinney's superposition. Scott McKinney's superposition results in "snipping" a snark G' along the dotted line and leaving all edges and vertices below the dotted line untouched. Above the dotted line, the edges labeled 6 and 7 are respectively "replaced by two Petersen graphs" as shown in the diagram. From [Ko], the resulting graph H in the right hand side of Figure 2 is a snark.

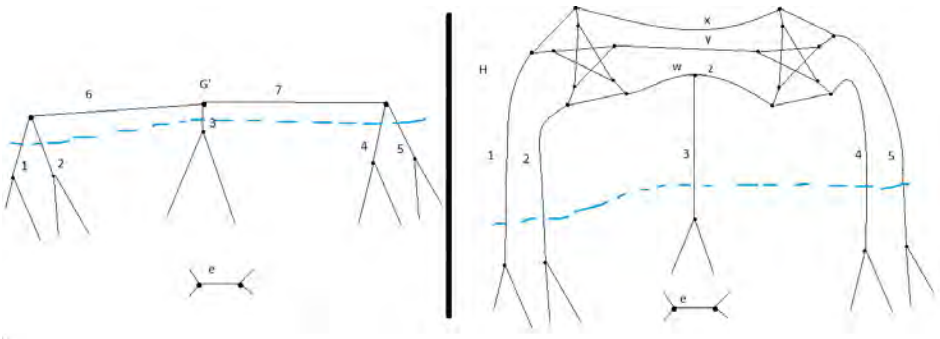


Figure 2: Scott McKinney's superposition.

The same edge e (somewhere below the dotted line) is indicated in both diagrams in Figure 2. Note that for any possible coloring of H_e, the edges 1 and 2

will never be assigned the same color. If they were, edges x, y, and w would all have to be assigned different colors (Parity Lemma, applied to edges $1, 2, x, y$, and w) and that would imply that there exists a coloring of the Petersen Graph. This is impossible because the Petersen Graph is a snark. Edges 4 and 5 can never be assigned the same color by similar reasoning. If they were, edges x, y, and z would all have to be assigned different colors (Parity Lemma, applied to edges $1, 2, x, y$, and z) and that also would imply that there exists a coloring of the Petersen Graph. Note then that for every possible coloring of H_e, either edge 1 or edge 2 must be assigned the same color as either edge 4 or edge 5 by the Pigeon Hole Principle. Let's call this "shared" color r. It follows that for any coloring of H_e, edge 3 must be colored r (Parity Lemma, applied to edges $1, 2, 3, 4$, and 5). It also follows that for any possible coloring of H_e there will be only two edges (out of edges 1, 2, 3, 4, and 5) not colored r and they will not share the same color (Parity Lemma).

Note that for any possible coloring of G'_e, the edges 1 and 2 will never be assigned the same color because they are adjacent edges. Edges 4 and 5 can never be assigned the same color by similar reasoning. Note then that for every possible coloring of G'_e either edge 1 or edge 2 must be assigned the same color as either edge 4 or edge 5 by the Pigeon Hole Principle. Again, let's call this "shared" color r. It follows that for any coloring of G'_e, edge 3 must be colored r. It also follows that for any possible coloring of G'_e there will be only 2 edges (among edges 1, 2, 3, 4, and 5) not colored r and they will not share the same color (Parity Lemma). Note once edges $1, 2, 3, 4$, and 5 are assigned colors, there is only one possible way to assign a color to edges 6 and 7. It follows that any coloring of H_e induces exactly one coloring of G'_e in which the colors of the "edges below the dotted line" are left unchanged.

In Figure 3, the snark H from (the right hand side of) Figure 2 is augmented to form a graph G which (from [Ko]) is also a snark. (In Figure 3, the portion "below the dotted line" is exactly the same as in Figure 2.) By an argument exactly analogous to the on eabove, any coloring of G_e induces exactly one coloring of G'_e (from Figure 2) with the colors below the dotted line unchanged. Our remaining task is to show that each possible coloring of G'_e leads to 17 possible edge colorings of G_e (i.e. H_e with certain new edges between the edges x,, y, and z, see Figure 3), again with no changes of colors to "edges below the dotted line." That would then imply that the number of colorings of G_e is exactly 17 times the number of colorings of G'_e; and hence by Definition 3.4 one would have $\psi(G, e) = 17 \cdot \psi(G', e)$.

4.2 First Proof of Lemma 4.3 for $p = 17$

Suppose the graph G'_e (see Figure 2) is colored. Without loss of generality, let us assume edge 3 is assigned the color c, edges 4 and 5 are colored b and c (in either order), and edges 1 and 2 are colored a and c (in either order). Let $\{x', y', z'\}$

be an example of notation showing x is assigned the color x', y is assigned the color y', and z is assigned the color z'. Note the three colors assigned to x, y, and z all have to add up to a (Parity Lemma, applied to edges 4, 5, x, y, and z). Also note z cannot be assigned the color c because edge 3 is colored c and is adjacent to edge z. Therefore, the 5 ways to complete the colorings of x, y, and z are $\{a, a, a\}$, $\{a, b, b\}$, $\{b, a, b\}$, $\{b, b, a\}$, and $\{c, c, a\}$.

Refer to the three "horizontal" edges x, y, and z in the snark G in Figure 3. Consider what happens when you start assigning colors to the edges of G_e where edge 3 joins the edge z and you "work to the right" along the "horizontal edges", as shown in Figure 4.

As previously noted, the only 5 ways to complete the colorings of x, y, and z are $\{a, a, a\}$, $\{a, b, b\}$, $\{b, a, b\}$, $\{b, b, a\}$, and $\{c, c, a\}$ (Parity Lemma). There are only 17 possible ways to assign colors to x, y, z, $x1$, $y1$, $z1$, $x2$, $y2$, $z2$, $x3$, $y3$, and $z3$ (see Figure 5). Of course, in each of these 17 color patterns, the colors of the five "unlabeled new edges" between x, y, z, and $x3$, $y3$, $z3$ in Figures 3 and 4 will be uniquely determined. Note that the notations z and $z1$ refer to the same edge, but it will be convenient to keep both notations. The same applies to edges $x2$ and $x3$.

It is important to note, for each of these 17 color patterns:

(i) The colors assigned to $x3$, $y3$, and $z3$ all must add up to a (Parity Lemma, applied to edges x, y, z, $x3$, $y3$, and $z3$).

(ii) Every possible coloring of the edges x, y, z, $x1$, $y1$, $z1$, $x2$, $y2$, $z2$, $x3$, $y3$, and $z3$ leads to exactly one possible coloring of the remaining edges in the "ripped" Petersen graph attached to edges $x3$, $y3$, and $z3$ (Lemma 2.5).

(iii) Every possible coloring of x, y, and z corresponds to exactly 1 way to color w.

(iv) The colors assigned to x, y, and w in Figure 3 all add up to b (by simple arithmetic in $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ for the colors of the edges 3, w, x, y, z).

(v) Every possible coloring of the edges x, y, and w leads to exactly one possible coloring of the remaining edges in the "ripped" Petersen graph attached to edges x, y, and w (Lemma 2.5).

From all the comments so far, it follows that every coloring of G'_e (for the snark G' and the edge e in the left side of Figure 2) induces exactly 17 colorings of G_e (for the snark G in Figure 3). That completes the proof of Lemma 4.3 for $p = 17$.

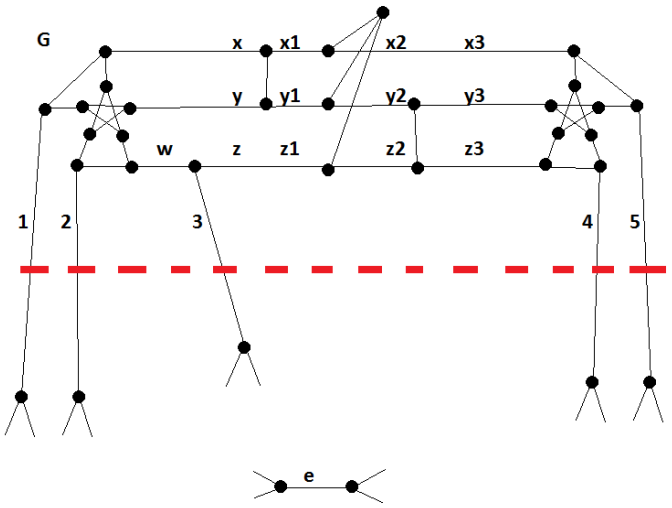


Figure 3: Scott McKinney's superposition with additional edges.

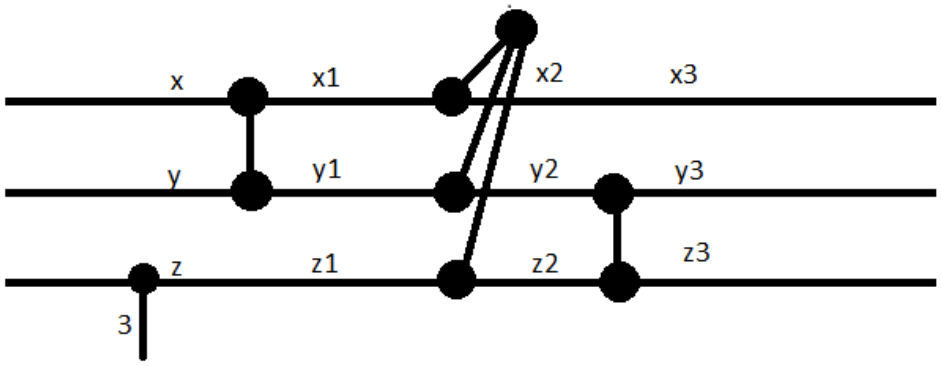


Figure 4: Close up of additional edges to produce 17 times as many colorings.

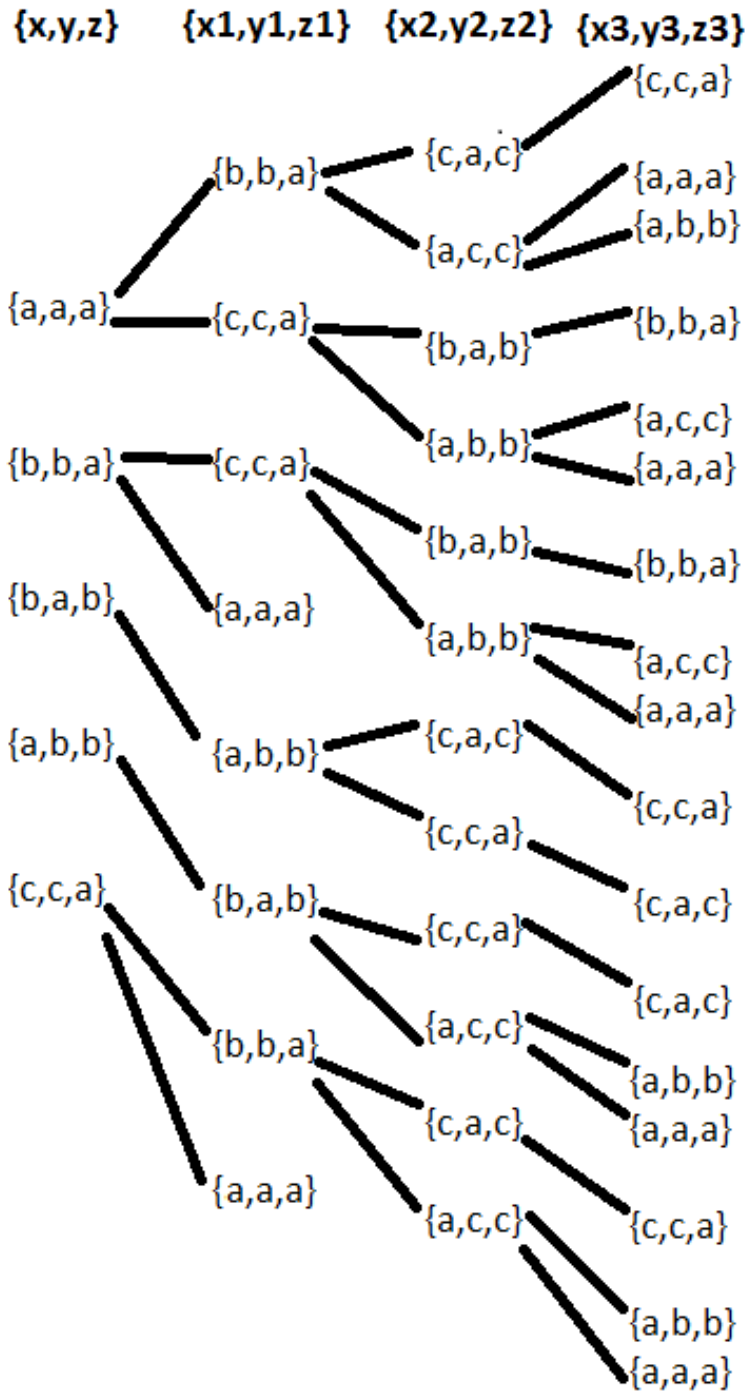


Figure 5: Chart of the 17 different ways you can assign colors to new edges.

4.3 Second Proof of Lemma 4.3 for $p = 17$

The construction shown in figures 3 and 4 can be modified to find Kászonyi numbers with all of the prime values in set $\mathscr{S}$. Described below is a method of "charting" these snarks minus an edge, to determine the number of edge-colorings of a graph with any array of added edges.

Consider the graph G_e where G and e are as in McKinney's superposition in Figure 2 (Note that the graph in figure 2 is called H, but to avoid ambiguities later, we will call it G). Consider the edges x, y, z (ref. figure 2). Again, we follow the stipulations in the first paragraph of Theorem 4.2, and thus again all of the colorings of these edges are {a,a,a}, {a,b,b}, {b,a,b}, {b,b,a}, and {c,c,a}. Thus these colorings of x, y, and z form four subsets:

i) The set containing the edge-colorings where $x = y = z$ is denoted S. The letter S stands for "same," because all x, y, and z are the same color. (Although x, y, and z are *edges*, for convenience we shall use equations such as $x = y = z$ to mean that these edges all have the same *color*. This "double use" of the symbols x, y, and z should be clear from the context.)
ii) The set where $x = y \neq z$ is denoted B. B stands for "bottom", because the bottom edge of the three-edge set has a *different* color from the top two edges.
iii) The set where $x = z \neq y$ is denoted M. M stands for "middle," because the middle edge is colored *differently* from the bottom and top edges.
iv) The set where $y = z \neq x$ is denoted T. T stands for top, because the top edge is colored *differently* from the bottom two edges.
Thus, $\{a, a, a\} \in S, \{a, b, b\} \in T, \{b, a, b\} \in M$, and $\{b, b, a\}, \{c, c, a\} \in B$.

Similarly, there are three edges that can be added to McKinney's graph, which are denoted B', M', and T', and are shown in Figure 6. B' denotes an edge that lies between the top two edges, so that the bottom edge remains unchanged. M' denotes an edge connecting the bottom and top edges, such that the middle edge in unchanged. T' denotes an edge connecting the bottom two edges, such that the top edge is unchanged.

In addition to these edges, there is a construction (also shown in figure 6) that can be inserted which is denoted B_A, standing for "banana" due to its resemblance of the luscious tropical fruit. This construction places a vertex in G, and connects each edge of the three-edge set to this vertex.

Note: The edge T' cannot be repeated consecutively, for that would create a 4-cycle, or "square." The same comment applies to each of the edges M' and B'. The B_A construction can be repeated consecutively, because that will not create squares.

The number of colorings of a graph will change depending on the number and order of added edges. We count the number of colorings by keeping track of the sets S, B, M, and T mentioned above. Note the following patterns:

1) Suppose we insert into a graph G an edge T'. For each color pattern

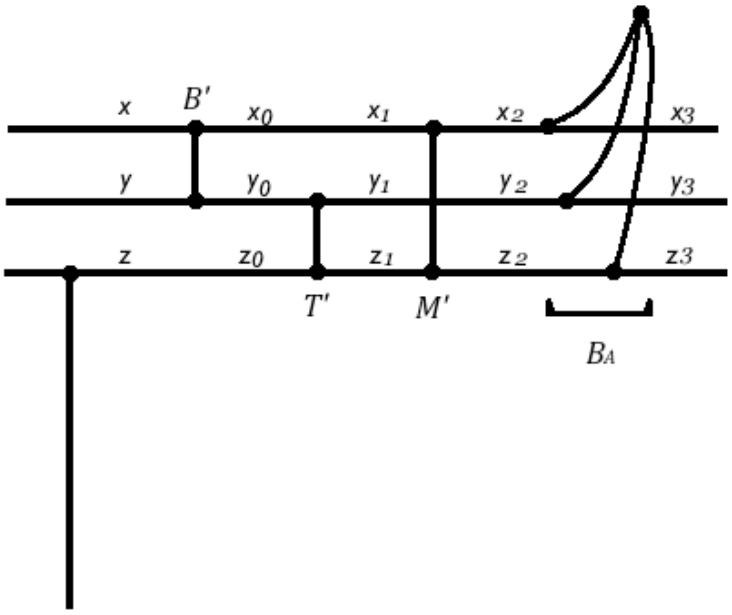


Figure 6: Possible additions to McKinney's graph

of $\{x_0, y_0, z_0\}$ in S, (i.e $x_0 = y_0 = z_0$), there are two possible color patterns of $\{x_1, y_1, z_1\}$ (arising, respectively, from the two possible choices of color to assign to the inserted edge T'.) They both satisfy $x_1 \neq y_1 = z_1$. So, for each coloring of $\{x_0, y_0, z_0\}$ in S there are two new colorings of $\{x_1, y_1, z_1\}$ in T. An equivalent comment applies for added edges M' and B', with sets M and B respectively.

Note: It should be kept in mind that the colors of the edges $x_0, y_0, z_0, x_1, y_1, z_1$ uniquely imply one particular color for the edge T' itself. Analogous comments apply to edges B' and M' and the three edges of B_A.

2) When T' is inserted, for each coloring of $\{x_0, y_0, z_0\}$ in the set T there are two resulting colorings of $\{x_1, y_1, x_1\}$, (because there are two possible choices of color to assign T'). Observe that one resulting color pattern will be in set T, while the other color pattern will be in set S. An equivalent argument can be made by replacing T and T' with either B and B' or M and M' respectively.

3) When edge T' is inserted, the number of colorings in $B \cup M$ does not change, because the coloring of T' is forced. Note that when T' is added, each coloring of $\{x_0, y_0, z_0\}$ in B will induce a color pattern of $\{x_1, y_1, z_1\}$ in M, while each coloring of $\{x_0, y_0, z_0\}$ in M will induce a color pattern of $\{x_1, y_1, z_1\}$ in B.

That is, with the addition of T', the numbers of colorings in M and B switches. The same argument can be made for an edge B' with sets M and T, or for an edge M' with sets T and B respectively.

4) Finally, we can see the effects of the addition of a B_A construction, as seen in figure 6. Because the three edges of a banana are connected to one vertex, they must each have a different color. Therefore, an edge coloring in set S will share a color with one edge of the banana, but there lies a contradiction in that two adjacent edges cannot have the same color. So (in figure 6) there are no possible color patterns in set S for edges $\{x_2, y_2, z_2\}$ that can extend to any color pattern for $\{x_3, y_3, z_3\}$.

The remaining color patterns of $\{x_2, y_2, z_2\}$ have two edges with the same color, and one edge with a different color. The edge of the B_A with equal color to the two same edges of $\{x_2, y_2, z_2\}$ must connect to the different edge (to avoid contradiction). The two remaining edges of the B_A can be colored with the two remaining colors in any order order. This leads to two colorings of the B_A, meaning that there are two resulting colorings in $\{x_3, y_3, z_3\}$ for each coloring of $\{x_2, y_2, z_2\}$ in B, M, or T. Note that for each coloring of $\{x_2, y_2, z_2\}$ in B, one of the two resulting colorings of $\{x_3, y_3, z_3\}$ is in T, and the other is in M. The same holds true for colorings M B and T, and for T M and B respectively. Thus, we end up with $B_{new} = T_{old} + M_{old}$, $T_{new} = B_{old} + M_{old}$, $M_{new} = B_{old} + T_{old}$, where the initial colorings are on the right side of the equations, and the new colorings are on the left.

Let us express the above observations in terms of an algorithm. Let the edges x, y, z also be denoted x_0, y_0, z_0. For each $n \geq 0$, let S_n (resp. T_n, resp. M_n resp. B_n) denote the number of colorings of the edges $x_0, y_0, z_0, x_1, y_1, z_1, \cdots, x_n, y_n, z_n$ and (if $n \geq 1$) all edges T', M', B', B_A "between" x_0, y_0, z_0 and x_n, y_n, z_n such that i) the color pattern of $\{x_0, y_0, z_0\}$ is one of $\{a, a, a\}$, $\{a, b, b\}$, $\{b, a, b\}$, $\{b, b, a\}$ or $\{c, c, a\}$ and ii) the color pattern of $\{x_n, y_n, z_n\}$ belongs to the set S (resp. T, resp. M resp. B).

From the observations above, we can "count" colorings in the following manner recursively: When we add an edge T',

$$\begin{aligned} T_{n+1} &= 2S_n + T_n \\ S_{n+1} &= T_n \\ B_{n+1} &= M_n \\ M_{n+1} &= B_n \end{aligned}$$

By adding the values on the left side of the equation, we obtain the total number of colorings assigned so far (starting with edges x, y, z and "working to the right", up to and including, say, the edges $x_{n+1}, y_{n+1}, z_{n+1}$).

Below are the equations for "counting" the colorings for the other added edges.
When B' is added:

$$\begin{aligned} B_{n+1} &= 2S_n + B_n \\ S_{n+1} &= B_n \end{aligned}$$

$$T_{n+1} = M_n$$
$$M_{n+1} = T_n$$

When M' is added:

$$M_{n+1} = 2S_n + M_n$$
$$S_{n+1} = M_n$$
$$T_{n+1} = B_n$$
$$B_{n+1} = T_n$$

When a B_A construction is added:

$$S_{n+1} = 0$$
$$T_{n+1} = B_n + M_n$$
$$B_{n+1} = T_n + M_n$$
$$M_{n+1} = B_n + T_n$$

Now, we can use this algorithm to show (as in our previous proof) that any coloring of G_e (where G' and e are as in Figure 2) induces exactly 17 colorings of G_e (where G is as in Figure 3).

Below is a chart that displays the algorithm listed above. The letters above each column denote whether T', B', M' or B_A were added to G. The letters before each row denote the number of colorings of the type S, T, B, M. The final number in each column denotes the total number of colorings of G "so far" (starting with edges x, y, z and "working to the right").

	5C	B'	BA'	T'	
S	1	2	0	5	
T	1	1	5	5	
M	1	1	5	2	
B	2	4	2	5	
Total:	5	8	12	17	

Figure 7: Possible additions to McKinney's graph

Observe that this chart gives the 17 possible color patterns for edges $\{x, y, z\}$ through edges $\{x_3, y_3, z_3\}$ in figures 3 and 4. Also, notice that the notes $(i) - (v)$ at the end of section 4.2 apply here as well. Thus, this chart in Figure 7 gives in "coded form" (again) the proof of lemma 4.3 for $p = 17$.

4.4 Extension of 4.3: Proof of Lemma 4.3 for any $p \in \mathscr{S}$.

In this section, we will use the same methods as used in section 4.3 to show that Lemma 4.3 holds for any $p \in \mathscr{S}$. (Recall from the comments after the statement of Lemma 4.3 that the prime numbers $p = 2, 3, 5, 7$ were already covered.)

In the previous section, we proved Lemma 4.3 for $p = 17$ through the chart in figure 7. Observe that the chart in figure 8 is an extension of the chart in

figure 7. The figure illustrates that multiple primes can arise from one chart. This particular chart yields in "coded form" a proof of the Lemma 4.3 for four primes: 17, 31, 127, and 257. (For example, for $p = 31$ we use precisely the sequence of operations B', B_A, T', B_A, T', just as we used B', B_A, T' for $p = 17$ in section 4.3.)

	5C	B'	BA'	T'	BA'	T'	BA'	T'	BA'	T'	BA'	T'	BA'	T'
S	1	2	0	5	0	7	0	17	0	31	0	65	0	127
T	1	1	5	5	7	7	17	17	31	31	65	65	127	127
M	1	1	5	2	10	7	17	14	34	31	65	62	130	130
B	2	4	2	5	7	10	14	17	31	34	62	65	127	127
Total:	5	8	12	17	24	31	48	65	96	127	192	257	384	511

Figure 8: Chart for $p = 17, 31, 127, 257$

Each prime $p \in \mathscr{S}$ such that $p \geq 11$, exists in at least one chart on the next four pages. To prove Lemma 4.3 for any given prime number $p \in \mathscr{S}$ such that $p \geq 11$, apply an analog of the proof given in section 4.3 for $p = 17$.

Chart of Primes															
2	3	5	7	11	13	17	19	23	29	31	37	41	43	47
53	59	61	67	71	73	79	83	89	97	101	103	107	109	113
127	131	137	139	149	151	157	163	167	173	179	181	191	193	197
199	211	223	227	229	233	239	241	251	257	263	269	271	277	281
283	293	307	311	313	317	331	337	347	349	353	359	367	373	379

Figure 9: List of primes.

	5C	M'	B'	M'	B'	M'	B'	M'	B'	M'	B'	M'	B'	M'
S	1	1	1	2	3	3	4	7	9	10	15	23	28	35
T	1	2	3	3	4	7	9	10	15	23	28	35	53	74
M	1	3	2	4	3	9	7	15	10	28	23	53	35	91
B	2	1	3	3	7	4	10	9	23	15	35	28	74	53
Total:	5	7	9	12	17	23	30	41	57	76	101	139	190	253

	5C	B'	BA'	T'	BA'	T'	BA'	T'	BA'	T'	BA'	T'	BA'	T'
S	1	2	0	5	0	7	0	17	0	31	0	65	0	127
T	1	1	5	5	7	7	17	17	31	31	65	65	127	127
M	1	1	5	2	10	7	17	14	34	31	65	62	130	130
B	2	4	2	5	7	10	14	17	31	34	62	65	127	127
Total:	5	8	12	17	24	31	48	65	96	127	192	257	384	511

	5C	B'	BA'	T'	BA'	M'	BA'	B'	T'	BA'	M'	T'	M'	
S	1	2	0	5	0	10	0	17	14	0	65	65	34	
T	1	1	5	5	7	7	17	14	48	34	65	195	65	
M	1	1	5	2	10	10	14	17	17	65	65	34	164	
B	2	4	2	5	7	7	17	17	17	65	34	65	195	
Total:	5	8	12	17	24	34	48	65	96	164	229	359	458	0

	5C	B'	T'	M'	T'	M'	T'	M'	BA'					
S	1	2	1	4	1	5	6	9	0					
T	1	1	5	1	9	6	16	7	37					
M	1	1	4	6	5	7	9	21	23					
B	2	4	1	5	6	9	7	16	28					
Total:	5	8	11	16	21	27	38	53	88	0	0	0	0	0

	5C	T'	B'	M'	T'	B'	M'	T'	B'	M'	T'			
S	1	1	1	3	3	5	9	11	19	29	41			
T	1	3	2	3	9	2	11	29	2	41	99			
M	1	2	3	5	2	9	19	2	29	67	2			
B	2	1	3	2	5	11	2	19	41	2	67			
Total:	5	7	9	13	19	27	41	61	91	139	209	0	0	0

	5C	T'	B'	M'	T'	BA'	B'	M'
S	1	1	1	3	3	0	11	7
T	1	3	2	3	9	7	14	11
M	1	2	3	5	2	14	7	29
B	2	1	3	2	5	11	11	14
Total:	5	7	9	13	19	32	43	61

	5C	M'	T'	B'	BA'	B'	BA'	B'
S	1	1	2	3	0	5	0	19
T	1	2	4	1	11	8	16	13
M	1	3	1	4	8	11	13	16
B	2	1	3	7	5	5	19	19
Total:	5	7	10	15	24	29	48	67

	5C	B'	M'	B'	T'	M'	B	T'	M'
S	1	2	1	1	5	3	7	13	13
T	1	1	4	5	7	4	13	27	4
M	1	1	5	4	3	13	4	13	39
B	2	4	1	3	4	7	13	4	27
Total:	5	8	11	13	19	27	37	57	83

	5C	B'	T'	B'	BA'	M'	T'	BA'	B'
S	1	2	1	1	0	7	9	0	31
T	1	1	5	4	8	9	23	15	30
M	1	1	4	5	7	7	8	30	15
B	2	4	1	3	9	8	7	31	31
Total:	5	8	11	13	24	31	47	76	107

	5C	B'	T'	M'	B'	BA'	M'		
S	1	2	1	4	5	0	19		
T	1	1	5	1	6	14	14		
M	1	1	4	6	1	19	19		
B	2	4	1	5	13	7	7		
Total:	5	8	11	16	25	40	59	0	0

	5C	B'	T'	M'	B'	T'	BA'	B'
S	1	2	1	4	5	6	0	29
T	1	1	5	1	6	16	14	17
M	1	1	4	6	1	13	17	14
B	2	4	1	5	13	1	29	29
Total:	5	8	11	16	25	36	60	89

	5C	M'	T'	M'	T'	M'	T'	M'	T'	M'	T'	M'	T'	M'
S	1	1	2	1	3	4	5	5	10	13	15	20	33	41
T	1	2	4	3	5	5	13	10	20	15	41	33	73	50
M	1	3	1	5	4	10	5	15	13	33	20	50	41	107
B	2	1	3	4	5	5	10	13	15	20	33	41	50	73
Total:	5	7	10	13	17	24	33	43	58	81	109	144	197	271

	5C	M'	T'	M'	T'	M'	T'	BA'	T'
S	1	1	2	1	3	4	5	0	15
T	1	2	4	3	5	5	13	15	15
M	1	3	1	5	4	10	5	23	18
B	2	1	3	4	5	5	10	18	23
Total:	5	7	10	13	17	24	33	56	71

	5C	M'	T'	M'	T'	M'	T'	BA'	M'
S	1	1	2	1	3	4	5	0	23
T	1	2	4	3	5	5	13	15	18
M	1	3	1	5	4	10	5	23	23
B	2	1	3	4	5	5	10	18	15
Total:	5	7	10	13	17	24	33	56	79

	5C	B'	M'	B'	M'	BA'	B'	T'	B'
S	1	2	1	1	4	0	9	8	11
T	1	1	4	5	3	11	8	26	9
M	1	1	5	4	6	8	11	9	26
B	2	4	1	3	5	9	9	11	27
Total:	5	8	11	13	18	28	37	54	73

	5C	M'	B'	M'	T'	M'	B'	M'	T'	M'	B'
S	1	1	1	2	3	3	7	4	13	9	21
T	1	2	3	3	7	4	9	13	21	18	35
M	1	3	2	4	3	9	4	18	9	35	18
B	2	1	3	3	4	7	13	9	18	21	39
Total:	5	7	9	12	17	23	33	44	61	83	113

	5C	T'	B'	BA'	T'	B'	M'	BA'	T'	
S	1	1	1	0	6	5	6	0	21	
T	1	3	2	6	6	5	17	21	21	
M	1	2	3	5	5	6	16	22	33	
B	2	1	3	5	5	17	5	33	22	
Total:	5	7	9	16	22	33	44	76	97	0

	5C	M'	T'	B'	M'	T'	M'	T'	M'	B'	M'	B'
S	1	1	2	3	4	7	1	10	15	12	15	35
T	1	2	4	1	7	15	10	12	15	35	42	39
M	1	3	1	4	10	1	15	15	35	15	39	42
B	2	1	3	7	1	10	15	15	12	42	35	65
Total:	5	7	10	15	22	33	41	52	77	104	131	181

	5C	M'	B'	T'	M'	B'	BA'	B'	M'	T'	
S	1	1	1	3	3	5	0	11	13	11	
T	1	2	3	5	2	9	13	20	11	37	
M	1	3	2	3	9	2	20	13	35	20	
B	2	1	3	2	5	11	11	11	20	35	
Total:	5	7	9	13	19	27	44	55	79	103	0

	5C	M'	T'	B'	M'	T'	B'	M'	BA'	B'	
S	1	1	2	3	4	7	10	15	0	59	
T	1	2	4	1	7	15	1	24	36	25	
M	1	3	1	4	10	1	15	35	25	36	
B	2	1	3	7	1	10	24	1	59	59	
Total:	5	7	10	15	22	33	50	75	120	179	0

	5C	M'	T'	B'	M'	T'	B'	BA'	B'	T'	
S	1	1	2	3	4	7	10	0	16	25	
T	1	2	4	1	7	15	1	39	25	57	
M	1	3	1	4	10	1	15	25	39	16	
B	2	1	3	7	1	10	24	16	16	39	
Total:	5	7	10	15	22	33	50	80	96	137	0

	5C	M'	T'	B'	BA'	B'	T'	BA'	M'	T'
S	1	1	2	3	0	5	8	0	29	23
T	1	2	4	1	11	8	18	16	23	81
M	1	3	1	4	8	11	5	29	29	16
B	2	1	3	7	5	5	11	23	16	29
Total:	5	7	10	15	24	29	42	68	97	149

	5C	M'	T'	B'	M'	T'	B'	M'	T'	B'	M'
S	1	1	2	3	4	7	10	15	24	35	54
T	1	2	4	1	7	15	1	24	54	1	83
M	1	3	1	4	10	1	15	35	1	54	124
B	2	1	3	7	1	10	24	1	35	83	1
Total:	5	7	10	15	22	33	50	75	114	173	262

5 Side Results

5.1 Extension to Pentagons

Definition 5.1 Let G be a snark. A pentagon P is any cycle contained in G with exactly five edges.

Note that for any pentagon P of G, the numbers $\psi(G, e)$, $e \in P$ are equal (see Theorem 4.5 in [Br2]). Let the common value of $\psi(G, e)$, $e \in P$ be denoted $\psi(G, P)$.

Theorem 5.2 *Suppose n is a positive integer of the form:*

$$n = \prod_{p \in \mathscr{S}} p^{m(p)}$$

where for each $p \in \mathscr{S}$ (see Definition 4.1), $m(p)$ is a nonnegative integer, and $m(3) \neq 1$. Then there exists a snark G and a pentagon P of G such that $\psi(G, P) = n$.

The proof is like that of Theorem 4.2, starting with the Petersen graph and using an induction argument via an analog of Lemma 4.3, (e.g. with the edge e in Figures 2 and 3 replaced by a pentagon P that is "entirely below the dotted line"). The prime factors 2 and 5 are implicitly handled by [Br1, Theorem 2.2] and [McK, Theorem 5.3] respectively.

All prime numbers $p \in \mathscr{S}$ such that $p \geq 7$ are handled by the collection of charts at the end of Section 4. It is not yet known how to cover the prime factor 3 in the scheme (for pentagons). The numbers $9 = 3^2$ and $27 = 3^3$ are coverd in the collection of charts; and by induction all higher powers of 3 are covered. Hence (at least for now) the situation $m(3) \neq 1$ as in the statement of Theorem 5.2.

5.2 3 Petersen Graphs 11, 19

While in the last section we discussed snarks that contain two Petersen graphs, here we can observe a snark that contains three Petersen graphs. Observe this in the graphs displayed in Figures 10 and 11.

The snark (in Figure 11) is a superposition of three Petersen graphs and a "base snark" G' shown in Figure 10. The edges $7, 8$, and 9 in Figure 10 are each replaced by a Petersen graph, as shown in Figure 11, to get a new, bigger snark G. The point to be made here is that $\psi(G, e) = 11 \cdot \psi(G', e)$ for a given edge e of the "base snark" G' that is *not* involved in the Figure 11 replacement of three edges by Petersen graphs.

To begin, subtract edge e from G to obtain the graph G_e. To find $\psi(G, e)$, we must find the number of possible colorings for G_e. Observe that the edges u, v, w must each be a different color. Recall that edges 1 and 2 must be colored differently, as must be edges 3 and 4, and 5 and 6. Because the edges $1, 2, 3, 4, 5, 6$ form a minimal cut set, the sum of their colorings must equal zero (by the Parity

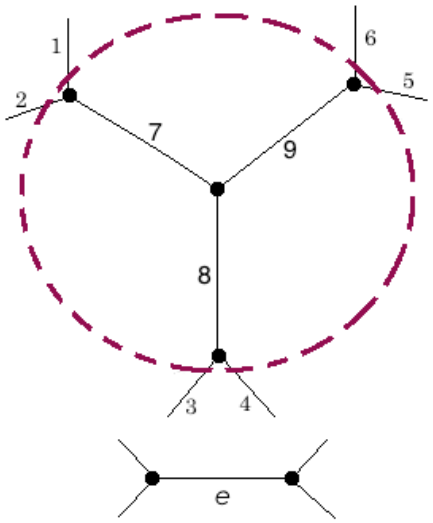


Figure 10: The base snark G'.

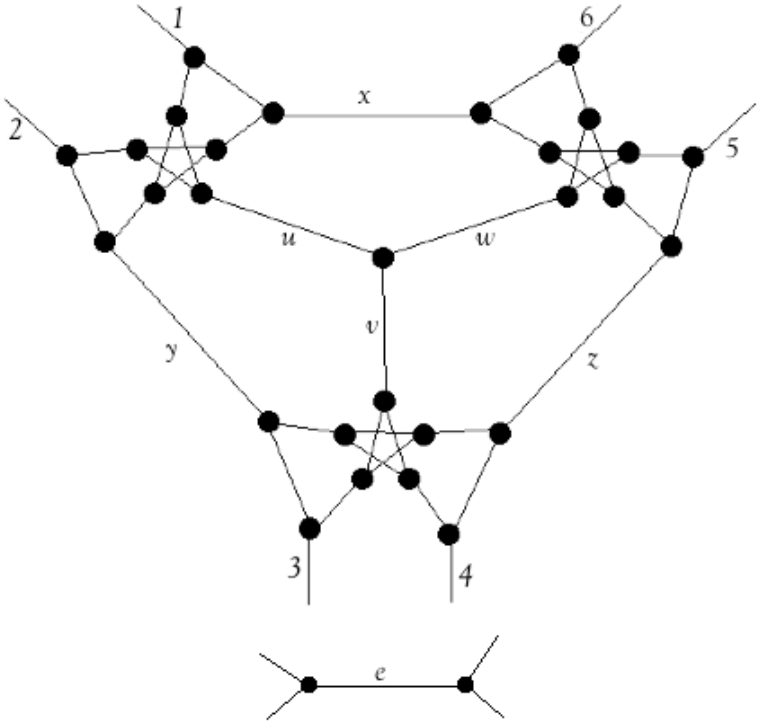


Figure 11: One way to combine 3 Petersen Graphs.

Lemma). Therefore, without loss of generality, edges 1 and 2 are colored a and b in either order, edges 3 and 4 are colored b and c in either order, and edges 5 and 6 are colored a and c in either order. Assigning the "forced" colors c, a, and b to edges $7, 8$, and 9 respectively, one sees that a given coloring of G_e induces a unique coloring of G'_e (with no changes outside the dotted circle).

Now suppose one has a coloring of G'_e. We would like to show that this will induce exactly 11 colorings of G_e (with no changes in colors "outside the dotted circle"). Note that the edges $\{1, 2, x, u, y\}$ form a (minimal) cut set, as do the edges $\{3, 4, y, v, z\}$, and $\{5, 6, x, w, z\}$. Thus in each of these sets, the colors of the edges must add to 0. Therefore, the colors of $\{x, u, y\}$ must add to c, the colors of $\{y, v, z\}$ must add to a, and the colors of $\{x, w, z\}$ must add to b.

Now consider the edges $\{x, y, z\}$. Because each of these edges has three color choices, the total number of permitted colorings is 27. These colorings are listed below:

x	y	z
a	a	a
a	a	b
a	a	c
a	b	a
a	b	b
a	b	c
a	c	a
a	c	b
a	c	c
b	a	a
b	a	b
b	a	c
b	b	a
b	b	b
b	b	c
b	c	a
b	c	b
b	c	c
c	a	a
c	a	b
c	a	c
c	b	a
c	b	b
c	b	c
c	c	a
c	c	b
c	c	c

Recall that the colors of $\{x, u, y\}$ must add to c, the colors of $\{y, v, z\}$ must add to a, and the colors of $\{x, w, z\}$ must add to b. Therefore, each set of edges can be assigned at most two distinct colors, and one of these colors must be

equal to their sum. So, of the edges $\{x, u, y\}$, if u is colored c, then x and y share a color. If u is not colored c, then one of x, y must be colored c. The same holds for edges $\{y, v, z\}$ and the coloring a, and the edges $\{x, w, z\}$ with the coloring b.

Following these facts, we conclude that x and y cannot be colored with b and a in either order, y and z cannot be colored b and c in either order, and x and z cannot be colored with a and c, again in either order. Therefore, we can eliminate the colorings with these properties from our list of choices above.

Below is the new list of possible colorings for edges x, y, z:

x	y	z
a	a	a
b	b	b
c	c	c
a	a	b
a	c	a
b	b	a
b	c	c
c	a	c
c	b	b
b	c	a
c	a	b

Observe that there are 11 possible colorings listed above. Each of these colorings forces exactly one coloring of the edges u, v, w, and then, (by Lemma 2.5) also forces exactly one coloring of the remaining edges in the three "ripped Petersen graphs." Thus, any coloring of G'_e induces exactly 11 colorings of G_e. It now follows that the number of colorings of G_e is exactly 11 times the number of colorings of G'_e, and hence $\psi(G, e) = 11 \cdot \psi(G', e)$.

A few additions to the graph G (in Figure 11) gives us a "new snark G" with $\psi(G, e) = 19 \cdot \psi(G', e)$. This graph is obtained by adding two new edges to G: one that connects edges x and u, and to the right of that, one that connects x and w. The edge we remove is still edge e.

5.3 "Legalized" Squares

We also looked at what happens to the Kászonyi number when you add "squares" into the edges x, y, and z. Even though squares are "illegal", if you attach a Petersen graph to a square (see Figure 12) the square is no longer illegal.

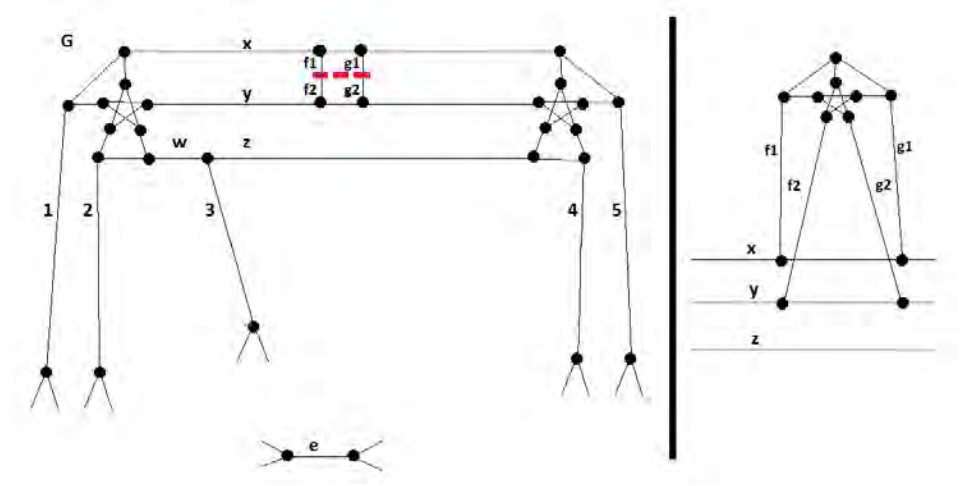


Figure 12: Example of making squares "legal"

Note that for any possible coloring of G_e with the attached Petersen graph (see Figure 12), $f1$ and $f2$ will always have to have to have the same edge color and $g1$ and $g2$ will also have to have the same edge color.

Proof Assume that an edge coloring of G_e exists where $f1$ and $f2$ do not have the same edge color and $g1$ and $g2$ also do not have the same edge color. Note the colors assigned to $f1$, $f2$, $g1$, and $g2$ all have to add to zero (Parity Lemma applied to edges $f1$, $f2$, $g1$, and $g2$). The only two ways for four edge colors to add up to 0 is to have all four edges colored the same color or have two sets of two edges colored the same color (Parity Lemma). Therefore, because $f1$ and $f2$ must have different colors, the two colors used to color $f1$ and $f2$ are also used to color $g1$ and $g2$. That would induce a coloring of the Petersen graph (see Figure 13), which is impossible. Therefore, any coloring of this new "snipped" Petersen graph would have to assign $f1$ and $f2$ the same color and $g1$ and $g2$ the same color. □

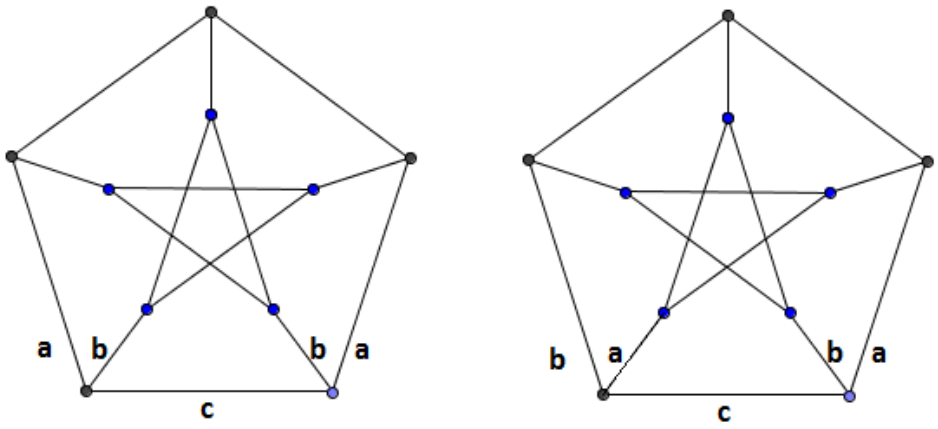


Figure 13: Example of an induced coloring of the Petersen graph (contradiction).

Note that for any possible coloring of $f1$, $f2$, $g1$, and $g2$ there will always be exactly two ways to fill in the remaining colors of the "snipped" Petersen graph in Figure 12. Therefore, for every snipped Petersen graph that needs to be "added" on to make squares legal, the total number of colorings of G_e will increase by a factor of 2 (by Theorem 3.5 and [Br2, Theorem 3.3 (C)(3)]). Therefore, unlike before, we can now apply T' repeatedly to the edges x, y, and z (as long as we realize we will have factors of 2 added into our new total multiple increase in the Kászonyi number of our original snark). When we do this, we get out a family of new snarks that have Kászonyi numbers with prime factors of the form $2^m + 3$ where $m \epsilon \mathbb{Z}$ (see Figure 14).

	5C	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'	T'
S	1	1	3	5	11	21	43	85	171	341	683	1365	2731	5461	
T	1	3	5	11	21	43	85	171	341	683	1365	2731	5461	10923	
M	1	2	1	2	1	2	1	2	1	2	1	2	1	2	
B	2	1	2	1	2	1	2	1	2	1	2	1	2	1	
Total:	5	7	11	19	35	67	131	259	515	1027	2051	4099	8195	16387	

Figure 14: Example of Multiple T' edges in a row.

5.4 Adding T' edges to both sides of Edge 3

Let w be a positive integer. Consider what happens when you add w T' edges to both sides of edge 3 of snark H (see Figure 2). Note any squares that are created can be "legalized" by attaching a certain snipped Petersen graph (see above section). For now with temporary inaccuracy, we will ignore the ending power of 2 that is caused by "legalizing" squares. Let's call the resulting snark Hw. McKinney previously proved that for every coloring of the original snark

G' (see Figure 2), there are only 5 ways to color the edges x, y, and z of snark H. For the same reasons, for every coloring of the original snark G' there are only 5 ways to color the edges s, t, and v (see Figure 15) of Hw. Note that each coloring of s, t, and v leads directly to a coloring of u.

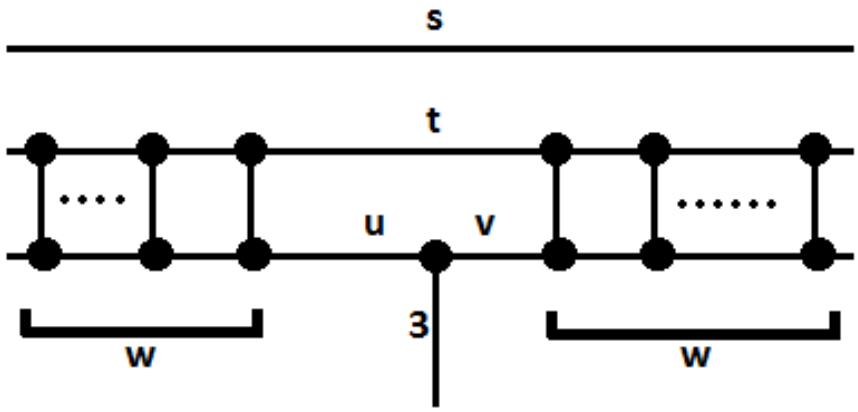


Figure 15: Close up near edge 3 of "illegal snark" Hw

Suppose the graph G'_e (see Figure 2) is colored. Without loss of generality, let us assume edge 3 is assigned the color c, edges 4 and 5 are colored b and c (in either order), and edges 1 and 2 are colored a and c (in either order). Let $\{s', t', v'\}$ be an example of notation showing s is assigned the color s', t is assigned the color t', and v is assigned the color v'. Note the three colors assigned to s, t, and v all have to add up to a (Parity Lemma, applied to edges 4, 5, s, t, and v). Also note v cannot be assigned the color c because edge 3 is colored c and is adjacent to edge v. Therefore, the 5 ways to complete the colorings of s, t, and v are $\{a, a, a\}$, $\{a, b, b\}$, $\{b, a, b\}$, $\{b, b, a\}$, and $\{c, c, a\}$.

For the coloring $\{a, a, a\}$, there are 2^w ways to finish coloring the T' edges on the right side of edge 3 (there will be 2 ways to color each T' edge to the right of edge 3 due to t and v having the same color). Note that when s, t, and u are colored $\{a, a, a\}$, edge u (see Figure 15) must be colored b (remember, edge 3 is colored c). Since edges t and u have different colors, there is only one way to complete the coloring of the additional T' edges on the left side of edge 3. Therefore, the $\{a, a, a\}$ coloring of G'_e will result in $1 \cdot 2^w$ colorings of Hw_e.

For the coloring $\{a, b, b\}$, there are 2^w ways to finish coloring the T' edges on the right side of edge 3 (there will be 2 ways to color each T' edge to the right of edge 3 due to t and v having the same color). Note that when s, t,

and u are colored $\{a, b, b\}$, edge u (see Figure 15) must be colored a (remember, edge 3 is colored c). Since edges t and u have different colors, there is only one way to complete the coloring of the additional T' edges on the left side of edge 3. Therefore, the $\{a, b, b\}$ coloring of G'_e will result in $1 \cdot 2^w$ colorings of Hw_e.

For the coloring $\{b, a, b\}$, there is one way to finish coloring the T' edges on the right side of edge 3 due to t and v having different colors. Note that when s, t, and u are colored $\{b, a, b\}$, edge u (see Figure 15) must be colored a (remember, edge 3 is colored c). Since edges t and u have the same color, there are 2^w ways to complete the coloring of the additional T' edges on the left side of edge 3. Therefore, the $\{b, a, b\}$ coloring of G'_e will result in $1 \cdot 2^w$ colorings of Hw_e.

For the coloring $\{b, b, a\}$, there is one way to finish coloring the T' edges on the right side of edge 3 due to t and v having different colors. Note that when s, t, and u are colored $\{b, b, a\}$, edge u (see Figure 15) must be colored b (remember, edge 3 is colored c). Since edges t and u have the same color, there are 2^w ways to complete the coloring of the additional T' edges on the left side of edge 3. Therefore, the $\{b, b, a\}$ coloring of G'_e will result in $1 \cdot 2^w$ colorings of Hw_e.

For the coloring $\{c, c, a\}$, there is one way to finish coloring the T' edges on the right side of edge 3 due to t and v having different colors. Note that when s, t, and u are colored $\{c, c, a\}$, edge u (see Figure 15) must be colored b (remember, edge 3 is colored c). Since edges t and u have different colors, there is only one way to complete the coloring of the additional T' edges on the left side of edge 3. Therefore, the $\{c, c, a\}$ coloring of G'_e will result in 1 coloring of Hw_e.

Therefore, for every one coloring of G'_e there will be a total of $4 \cdot 2^w + 1$ colorings of Hw_e. Note that $4 \cdot 2^w + 1 = 2^{w+2} + 1$. Therefore (if squares are not "legalized"), $\psi(Hw, e) = [2^{w+2} + 1] \cdot \psi(G', e)$. However, remember we intentionally forgot the power of 2 caused by "legalizing" squares, so $\psi(Hw, e) =$ [some power of 2] $\cdot\, [2^{w+2} + 1] \cdot \psi(G', e)$.

Following the fact that $\psi(Hw, e) =$ [some power of 2] $\cdot [2^{w+2}+1] \cdot \psi(G', e)$ (for G' in Figure 2) and Theorem 3.5, we have the following theorem from letting the "base" snarks G' (in Figure 2) be the Petersen graph and employing arguments from Section 4:

Theorem 5.3 *For a given integer $m \geq 3$, there exist a snark G, an edge e of G, and a positive integer k such that $\psi(G, e) = 2^k \cdot [2^m + 1]$.*

5.5 Proof of an infinite family of prime factors

The following result demonstrates that there are infinitely many prime divisors of Kászonyi numbers.

Theorem 5.4 *There exist infinitely many prime numbers p such that the following holds: There exist a snark G, an edge e of G, and a positive integer k such that* $\psi(G, e) = k \cdot p$.

The proof below was pointed out by Richard Bradley; it is just a standard argument in elementary number theory.
Proof Let $S = \{$primes p: $\exists$ odd m such that $2^m + 1$ is a multiple of $p\}$. By Theorem 5.3, it suffices to show that the set S is infinite. For the sake of contradiction, suppose the set S is finite. Let the elements of S be denoted $\{p_1, p_2, ..., p_n\}$. For each $k \in \{1, 2, 3, ..., n\}$ let m_k be an *odd* integer such that $2^{m_k} + 1 \equiv 0 \bmod p_k$. For each $k \in \{1, \cdots, n\}$, the following holds:

i) $2^{m_k} \equiv -1 \bmod p_k$
ii) For any odd integer u, $2^{m_k \cdot u} = (2^{m_k})^u \equiv (-1)^u = -1 \bmod p_k$

Note that the integer $m_1 \cdot m_2 \cdot \ldots \cdot m_n$ is odd, as is the product of any subcollection of the m_k's. For each $k \in \{1, 2, \ldots, n\}$, $2^{m_1 \cdot m_2 \cdot \ldots \cdot m_n} \equiv -1 \bmod p_k$. Thus for each $k \in \{1, 2, \ldots, n\}$, $2^{m_1 \cdot m_2 \cdot \ldots \cdot m_n} + 1 \equiv 0 \bmod p_k$.

Let p be any prime factor of the number $2^{m_1 \cdot m_2 \cdot \ldots \cdot m_n + 2} + 1$. Now the integer $(m_1 \cdot m_2 \cdot \ldots \cdot m_n) + 2$ is odd. Hence, $p \in S$. Hence there exist $k \in \{1, \ldots, n\}$ such that $p = p_k$. Thus, $2^{m_1 \cdot \ldots \cdot m_n} + 1$ is a multiple of p (by the last sentence of the preceding paragraph). Hence $4 \cdot (2^{m_1 \cdot \ldots \cdot m_n} + 1) = 2^{(m_1 \cdot \ldots \cdot m_n) + 2} + 4$ is a multiple of p. Therefore p divides the difference $[2^{(m_1 \cdot \ldots \cdot m_n) + 2} + 4] - [2^{(m_1 \cdot \ldots \cdot m_n) + 2} + 1] = 3$. Hence $p = 3$. We have shown p is the only prime divisor of $2^{(m_1 \cdot \ldots \cdot m_n) + 2} + 1$.

Hence for some positive integer L, $2^{(m_1 \cdot \ldots \cdot m_n) + 2} + 1 = 3^L$. By a result in number thoery (see [McR]), the only solution of $2^a + 1 = 3^b$ for integers $a, b \geq 2$ is $2^3 + 1 = 3^2$. Hence $m_1 \cdot \ldots \cdot m_n + 2 = 3$, and $L = 2$. Thus, $m_1 \cdot \ldots \cdot m_n = 1$, and $m_k = 1$ for each $k \in \{1, \ldots, n\}$.

For each $k \in \{1, \ldots, n\}$, $2^1 + 1 \equiv 0 \bmod p_k$, i.e. $2^1 + 1$ is a multiple of p_k, i.e. 3 is a multiple of p_k. Hence each p_k can only be 3. That is, $S = \{3\}$. However, 11, for example, is in S, since $2^5 + 1 = 33 \equiv 0 \bmod 11$, (so 3 cannot be the only element of S). Thus, the set S is infinite after all, and Theorem 5.4 follows. □

6 Acknowledgements

First and foremost, we would like to thank professor Richard Bradley for his generosity. He wrote a survey paper to help us learn the background information for this problem, and he spent countless amounts of time and effort to ensure

that we had a good research experience. We would like to thank Scott McKinney for letting us build on his past results and work which made all of our results possible. We would like to thank the Indiana Mathematics REU program for hosting us and setting us up with a wonderful opportunity to learn new math. Lastly, we would like to thank the US National Science Foundation (NSF) for their sponsorship.

References

[Br1] R. Bradley: *On the Number of Colorings of a Snark Minus and Edge*, J Graph Theory 51 (2006), 251-259.

[Br2] R. Bradley: *Snarks from a Kászonyi Perspective: A Survey*, arXiv:1302.2655v1[math.CO] (2013).

[Ga] M. Gardner: *Mathematical games: Snarks, boojums, and other conjectures related to the four-color map theorem.* Sci. Amer. 234, No. 4, 126-130, 1976.

[Is] R. Isaacs: *Infinite families of nontrivial trivalent graphs which are not Tait colorable.* Amer. Math. Monthly 82 (1975) 221-239.

[Ka1] L. Kászonyi: *A construction of cubic graphs, containing orthogonal edges.* Ann. Univ. Sci. Budapest Eőtvős Sect. Math. 15 (1972) 81-87.

[Ka2] L. Kászonyi: *On the nonplanarity of some cubic graphs.* Ann. Univ. Sci. Budapest Eőtvős Sect. Math. 16 (1972) 123-131.

[Ka3] L. Kászonyi: *On the structure of coloring graphs.* Ann. Univ. Sci. Budapest Eőtvős Sect. Math. 16 (1973) 25-36.

[Ko] M. Kochol: *Snarks without small cycles.* J Combin. Theory Ser B 67 (1996) 34-47.

[McK] S. A. McKinney: *On the number of edge-3-colourings of a snipped snark.* arXiv:1304.5427v1[math.CO] (2013).

[McR] G. McRae: *Catalan's Conjecture:* $3^2, 2^3$ *are the only powers that differ by 1.* http://2000clicks.com/mathhelp/PuzzleUnsolvedCatalan.aspx (2013).

[Wi] R. Wilson: *Four Colors suffice.* Princeton University Press, Princeton (2002).

Interior Points of Strictly Convex C^2 Billiards are Generically Insecure

Tom Dauer

Abstract

A mathematical billiard is defined as a plane domain ("table") and a point mass that moves with constant speed inside the table in such a way that when the point mass hits the boundary, its angle of incidence equals its angle of reflection. Let x and y be points in a given table, possibly with $x = y$, either in the interior of the table or on the boundary. A blocking set for the pair (x, y) is a set of points in the table such that every billiard path from x to y passes through a point in the set. If a finite blocking set exists, the pair (x, y) is called secure; if not, it is called insecure. We show that given x and y, there exists a dense G_δ set in the space of strictly convex C^2 billiard tables with x and y in the interior for which the pair (x, y) is insecure. (A G_δ set is defined to be a countable intersection of open sets). In this sense, the pair (x, y) is insecure for a "generic" strictly convex C^2 billiard table with x and y in the interior. In 2009, Tabachnikov showed that if x and y are on the boundary of such a table, then the pair (x, y) is insecure; our result sheds light on the case in which x and y are in the interior.

1 Introduction

Consider a plane region ("table") bounded by a strictly convex C^2 closed curve $\sigma : S^1 \to \mathbb{R}^2$. A *billiard* is the dynamical system consisting of this table and a point mass inside the table that moves with constant speed in such a way that when it hits σ, its angle of incidence equals its angle of reflection (this is called the billiard reflection law). Let x and y be points in a given table M, either in the interior of the table or on the boundary. A *blocking set* for the pair (x, y) is a set of points in $M \setminus \{A, B\}$ such that every billiard path in M from x to y passes through a point in the set. If a finite blocking set exists, the pair (x, y) is called *secure*; if not, it is called *insecure*. A table is called secure if for each pair of points in the table, a finite blocking set exists; if not, it is called insecure. We call a point where a billiard path intersects the boundary a *vertex*.

In 2009, Tabachnikov showed that every compact plane billiard M bounded by a smooth curve (also known as a *Birkhoff billiard*, after G.D. Birkhoff, one of the founders of the study of dynamical systems) is insecure; see [5]. Tabachnikov considers a strictly convex arc $\gamma \subset \partial M$, possibly with $x = y$, which must exist since ∂M is smooth and closed. Let $A, B \in \gamma$. He considers the maximum length polygonal path L_n from A to B with all n vertices in γ, which is a

billiard path (see Lemma 3.1). He notes that if n is large, then L_n lies in a small neighborhood of γ; so if a finite set of blocking points for the pair (A, B) were to exist, it would only contain points on γ. Tabachnikov then uses the theory of interpolating Hamiltonians and some results on rational approximation to show that such a set cannot exist, which proves that the pair (A, B) is insecure.

Throughout the rest of this paper, let M be the closed region bounded by a strictly convex C^2 curve $\sigma : S^1 \to \mathbb{R}^2$. We consider points x and y in the interior of M. To prove that the pair (x, y) is insecure, it suffices to show that for any positive integer n, there exist n billiard paths from x to y that have no triple intersections except at x at y. A triple intersection is a point at the intersection of three of these paths. To see this, notice that if there were a finite blocking set S for the pair (x, y), then each point in S could block at most two billiard paths. Thus for $n > 2|S|$, the set S cannot block all billiard paths from x to y, which is a contradiction.

Let x and y be points in $\mathbb{R}^2$, possibly with $x = y$. Let $\mathcal{C}(x, y)$ be the set of strictly convex C^2 curves $\sigma : S^1 \to \mathbb{R}^2$ such that x and y are in the interior of the region enclosed by σ. We consider $\mathcal{C}(x, y)$ with the C^2 topology. For each $n \in \{1, 2, \ldots\}$, we will obtain a dense open subset $G_n(x, y)$ of $\mathcal{C}(x, y)$ such that for any $\sigma \in G_n(x, y)$, there exist n billiard paths from x to y with no triple intersections (except at x and y) in the table bounded by σ. Then we apply a corollary of the Baire category theorem to show that the intersection of these G_n's is dense. In this sense, the pair (x, y) is insecure for a table bounded by a "generic" curve in $\mathcal{C}(x, y)$. While Tabachnikov's result shows the insecurity of a specified pair of boundary points, we do not get any specific examples of insecure pairs of interior points from our argument here. We expect that our techniques will also show that for a generic $\sigma \in \mathcal{C}$ and a generic set of pairs of points (x, y) in the interior of the region enclosed by σ, the pair (x, y) is insecure for the table bounded by σ. We also expect to get a similar result for polygonal tables by using a reflection argument and the techniques used here.

While our methods do not allow us to draw a conclusion about security for any particular pair of points in a the interior of particular table, our result is in the spirit of results about ergodicity of billiard systems in polygons. It was shown in [3] that the directional billiard flow of a rational polygon is ergodic for almost every direction (but not all directions since the polygon is rational), and that the set of ergodic n-gons is residual in the sense of Baire category (we will define *residual* when we discuss the Baire category theorem in Section 3). This means that in the sense of Baire category there is a generic set of irrational ergodic polygons. However, no explicit examples of ergodic polygons were given until eleven years later; see [7]. There are still no known examples of irrational but non-ergodic polygons; see [2] for details.

Our results are analogous to insecurity results obtained for manifolds, where there is a similar definition for a blocking set, but which uses geodesics instead of billiard paths. Let M be a compact C^∞ manifold without boundary, of dimension at least two, and let $(x, y) \in M \times M$. It was shown in [1] that there exists a dense G_δ set of C^∞ Riemannian metrics g on M such that the pair

(x, y) is insecure. Moreover, according to [1], the set

$$\tilde{\mathcal{G}} = \{(x, y, g) \in M \times M \times \mathbb{G} : (x, y) \text{ is insecure in } g\}$$

contains the intersection of a countable collection of sets that are C^1-open and C^∞-dense in $M \times M \times \mathbb{G}$. This is analogous to a result we expect to obtain for plane billiards: for a generic (x, y, σ) with $\sigma : S^1 \to \mathbb{R}^2$ a strictly convex C^2 curve and x, y in the interior of the region enclosed by σ, the pair (x, y) is insecure. However, in [1], the Riemannian metric can be perturbed within the manifold M, while in our case the allowable perturbations are restricted to the boundary of the table M, which gives us less flexibility.

Work on security problems in polygonal billiard tables has been an active area of research for the past fifteen years. Monteil showed in [4] that there exists a rational polygonal billiard table (i.e. one for which all angles are rational multiples of π) that is insecure, contradicting some previous work in the area. It has also been shown (see [2] for extensive references) that the only secure regular polygons are equilateral triangles, squares, and regular hexagons. There are no known nontrivial examples of secure pairs of points in billiard tables with strictly convex C^2 boundary. A trivial example is the center of a circle and a point on the circle; this pair is trivially secure since the midpoint of the segment joining these two points blocks all billiard paths between them. One might expect that there are no other examples of secure pairs of points in billiard tables with strictly convex C^2 boundary, but this appears to be a very hard problem. Our contribution in this paper is showing that the set of insecure pairs in such tables is "large" in a topological sense.

2 Outline of Our Approach

Let M be the closed region bounded by a strictly convex C^2 curve $\sigma : S^1 \to \mathbb{R}^2$. Saying σ is strictly convex means that its curvature is strictly positive. We consider points x and y in the interior of M. As we explained above, to show that the pair (x, y) is insecure, it suffices to show that for every positive integer n, there exist n billiard paths from x to y that have no triple intersections except at x and y.

We proceed by induction to show that there for every positive integer n, there are n billiard paths from x to y that satisfy the following conditions:

1. There is no periodic path that uses only vertices from these n paths.
2. No two paths share a vertex.
3. These paths have no triple intersections except at x and y.
4. The points x and y are not conjugate along any of these paths. (See Definition 3.6.)

To prove this, we proceed by induction. The base case (one path from x to y) is trivial. Suppose that $n > 1$ is an integer and there are n paths satisfying

$(1)-(4)$. We wish to show that there exists an $(n+1)$st path from x to y that is distinct from the first n paths and satisfies conditions $(1)-(4)$.

Consider $\mathcal{P} = \{p_1, \ldots, p_k\}$, the vertices of the first n paths. Since there are (by condition (1) and the inductive hypothesis) no periodic paths using only vertices in $\mathcal{P}$, we can make a simple argument to show that by choosing a sufficiently large N, there exists a billiard path with N segments with at least one of its vertices, say p', not in $\mathcal{P}$. We call this our $(n+1)$st path, and we will modify the table and this path slightly so that all $n+1$ paths satisfy conditions $(1)-(4)$. We can make small perturbations of the table so that the boundary remains C^2 and strictly convex. When we refer to perturbing the table, we actually mean perturbing its boundary σ.

Consider families of rays from x to a small neighborhood of p' and from y to a small neighborhood of p' (see Proposition 5.3). Call the first family F_1 and the second F_2. (We think of rays as not only one directed line segment, but also all the reflections of that line segment off the boundary according to the billiard reflection law). We slightly change the curvature of σ in a small neighborhood of p' so that the families F_1 and F_2 do not focus at any vertex in $\mathcal{P}$ along the $(n+1)$st path (see the definition of focusing before Lemma 3.3), and so that x is not conjugate to y along the $(n+1)$st path. Therefore the $(n+1)$st path satisfies condition (4). We then show that there exists a ray in F_1 and a ray in F_2, neither of which is perpendicular to the boundary at any point or hit any of the vertices in $\mathcal{P}$, that we can "match up" by perturbing σ in a small neighborhood of p'. This gives us an $(n+1)$st path with no vertices in $\mathcal{P}$, so the collection of $n+1$ paths satisfies condition (2).

Now we slightly perturb the table finitely many times in small neighborhoods of the vertices of the $(n+1)$st path to avoid periodic paths that uses only vertices from the $n+1$ paths (see Corollary 5.7). Each of these perturbations changes the angle a segment on the $(n+1)$st path makes with the boundary at one of the new vertices, and this change is taken to be small enough that the altered vertices of the $(n+1)$st path still do not coincide with any of the vertices in $\mathcal{P}$. Now the $(n+1)$ paths satisfy condition (1).

Next, we slightly perturb the table finitely many times in small neighborhoods of the vertices of the $(n+1)$st path to eliminate triple intersections (except at x and y) that may have arisen from intersections of segments of the $(n+1)$st path with segments of the first n paths (see Corollary 5.8). There is sufficient flexibility in the way we do this perturbation to ensure that we introduce no new periodic paths (if our perturbation did introduce such a path, we could make our perturbation slightly smaller so that it did not). Now there are no triple intersections in the table except at x and y, so condition (3) is satisfied. This completes the outline of our inductive argument.

Our final step is to show that if $\tau : S^1 \to \mathbb{R}^2$ is a strictly convex C^2 curve bounding a table for which there exist n billiard paths from x to y satisfying $(1)-(4)$, then (2) and (3) remain true for sufficiently small C^2 perturbations of τ (see Lemma 5.11). We then have that for fixed n, there exists a dense open set of strictly convex C^2 boundary curves for which there exist n paths from x to y with no triple intersections except at x and y. We can then use a corollary of

the Baire category theorem to show that the G_δ set formed by the intersection of these dense open sets over all n is a dense set of strictly convex C^2 boundary curves with x and y in the interior for which the pair (x, y) is insecure.

3 Preliminary Results

The goal of this section is to prove a number of basic results that will be needed later. The results of this section are not new; we provide them here simply for the convenience of the reader.

Let M be as above, with strictly convex C^2 boundary σ, and let $A, B \in M$. Denote by L_n a polygonal line $AP_1 \dots P_nB$ of maximum length, where $P_i \in \partial M$ for all $i \in \{1, \dots n\}$. Then $(P_1, \dots P_n) \in \partial M \times \cdots \times \partial M$, which is a compact set, so such a maximum length line L_n exists. Note that all of the vertices of L_n must be distinct in order for L_n to have maximal length. For example, if $A = P_1$ then the path L_n' formed by moving P_1 slightly away from A on σ has length longer than L_n by the triangle inequality.

3.1 Some basic facts

Lemma 3.1 *L_n is a billiard trajectory.*

Proof We begin with the case $n = 2$. We assume σ parametrized at constant speed, with $\mathfrak{L}$ the total length of σ. Let $f(s) = \text{dist}(A, \sigma(S^1)) + \text{dist}(B, \sigma(S^1))$, where $s \in [0, \mathfrak{L}]$. Any maximum of f must be a local maximum, and since $s \in S^1$ (a compact set), f must attain a maximum on this interval. Thus, the maximum must occur at a critical point of f. We have

$$\begin{aligned}
\frac{d}{ds}\text{dist}(A, \sigma(s)) &= \frac{d}{ds}\left(\sqrt{(\sigma_1(s) - a_1)^2 + (\sigma_2(s) - a_2)^2}\right) \\
&= \frac{(\sigma_1(s) - a_1)\sigma_1'(s) + (\sigma_2(s) - a_2)\sigma_2'(s)}{\sqrt{(\sigma_1(s) - a_1)^2 + (\sigma_2(s) - a_2)^2}} \\
&= \frac{\vec{AP_1} \cdot \sigma'(s)}{\|\vec{AP_1}\|} \\
&= \frac{\|\vec{AP_1}\| \cdot \|\sigma'(s)\| \cos\theta_1}{\|\vec{AP_1}\|} \\
&= \cos\theta_1
\end{aligned}$$

And similarly $\frac{d}{ds}\text{dist}(B, \sigma(s)) = -\cos\theta_2$. Thus if $f'(s) = 0$ we have $0 = \cos\theta_1 - \cos\theta_2$, and hence $\theta_1 = \theta_2$ since $\cos x$ is $1 - 1$ for $x \in [-1, 1]$. Thus, AP_1B is a billiard path.

Now we consider T_n for $n > 2$. Suppose T_n is not a billiard path, so there exists $i \in \{1, \dots, n-1\}$ such that $\theta \neq \theta'$, where θ and θ' are the angles made by $P_{i-1}P_i$ and P_iP_{i+1}, respectively, with the tangent to γ at P_i. By the case above, this means that the arc length coordinate of P_i is not a critical point of

the function $g(s) := \text{dist}(P_{i-1}, P_i(s)) + \text{dist}(P_{i+1}, P_i(s))$. Thus the sum of the lengths of segments $P_{i-1}P_i(s)$ and $P_{i+1}P_i(s)$ is not maximized, so T_n is not a maximal trajectory, which is a contradiction to the definition of T_n (note that T_n must exist since each P_i has arc length coordinate in $[0, \mathfrak{L}]$, a compact set). Hence T_n is a billiard path for all $n \geq 1$. □

Assume $\sigma : S^1 \to \mathbb{R}^2$ is periodic, i.e. there exists an L such that $\sigma(s) = \sigma(s + L)$ for all $s \in S^1$.

Lemma 3.2 *The following are equivalent:*

1. *σ is strictly convex, i.e. $\kappa(s) > 0$ for all $s \in [0, L]$, where $\kappa(s)$ denotes the curvature of σ at s.*
2. *For each $s_0 \in [0, L]$, there exists $c > 0$ and a Cartesian coordinate system (u, v) such that $(u, v) = (0, 0)$ corresponds to $\sigma(s_0)$, $\sigma'(s_0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, and there is a strictly convex function $g : \mathbb{R} \to \mathbb{R}$ such that $v = g(u)$ and σ is the graph of g for $u \in (-c, c)$.*

Proof We define the unit tangent and unit normal vectors to σ at s by $\mathbf{T}(s) = \sigma'(s)/|\sigma'(s)|$ and $\mathbf{N}(s) = \mathbf{T}'(s)/|\mathbf{T}'(s)|$. Assume that $(\mathbf{T}, \mathbf{N})$ is positively oriented. We have $\mathbf{T}(s) \perp \mathbf{N}(s)$, and we define the curvature of γ at s by $\mathbf{T}'(s) = \kappa(s)\mathbf{N}(s)$. Write $\mathbf{T}(s) = (\sin\theta, \cos\theta)$ where $\theta \in [0, 2\pi)$; then $\mathbf{N}(s) = (-\sin\theta, \cos\theta)$. We have

$$\gamma''(s) = \mathbf{T}'(s) = \frac{d\theta}{ds}(-\sin\theta(s), \cos\theta(s)) = \frac{d\theta}{ds}\mathbf{N}(s)$$

so $\kappa(s) = \frac{d\theta}{ds}$. Since $\frac{d\theta}{ds} = g'(u) = \frac{du}{ds}$ and $\frac{du}{ds} > 0$, we see that $\kappa(s) > 0 \implies g'(u) > 0$. □

3.2 Families of lines and focusing

We now give some definitions and state some results on focusing; see [9] for details. An *oriented line* is a line in the plane along with a specified unit vector $\mathbf{v}$ parallel to line, giving it an orientation. Such a line is parametrized by $t \mapsto \mathbf{x} + t\mathbf{v}$ for some point $\mathbf{x}$ on l. A C^1 *family of oriented lines* $l(u)$, $u \in I$ for some open interval I, is parameterized by $l(u, t) = \gamma(u) + t\mathbf{v}(u)$, $t \in \mathbb{R}$, where $\gamma : \mathbb{R} \to \mathbb{R}^2$ and $\mathbf{v} : \mathbb{R} \to S^1 \subset \mathbb{R}^2$ are C^1 functions. The point $\gamma(u)$ and the vector $\mathbf{v}(u)$ are known as the *base point* and *direction vector*, respectively, of the line $l(u)$.

If $f(u) = -\langle \gamma'(u), \mathbf{v}'(u) \rangle / \langle \mathbf{v}'(u), \mathbf{v}'(u) \rangle$ (where $\mathbf{v}' \neq \mathbf{0}$), then $l(u, f(u))$ is called the *local envelope* of the family $l(u)$. Note that the line $l(u_0)$ is tangent to the envelope $f(u)$ at $u = u_0$. The point $F = l(u_0, f(u_0))$ is said to be the *focusing point* (in linear approximation) for the family $l(u)$ at $u = u_0$. The envelope can be thought of as the curve determined by the focusing points of $l(u)$; see Figure 1.

We now show that the choice of the curve of base points for a family of lines $l(u)$ does not affect its focusing points.

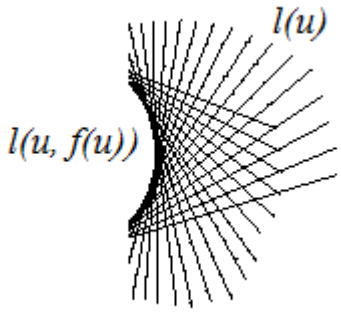


Figure 1: An envelope of lines

Lemma 3.3 *Suppose $l(u)$, $|u| < \delta$, is a C^1 family of oriented lines parameterized by using two different C^1 curves $\gamma(u)$ and $\tau(u)$ as base points. Then the focusing point F for l at $u = u_0$ is the same for both these parameterizations.*

Proof We may assume that $u_0 = 0$. Let $l(u,t) = \gamma(u) + t\mathbf{v}(u)$ and $j(u,t) = \tau(u) + t\mathbf{v}(u)$ be parameterizations of l. The family $l(u)$ focuses about $u = 0$ at the point

$$l\left(0, -\frac{\langle \gamma'(0), \mathbf{v}'(0)\rangle}{\langle \mathbf{v}'(0), \mathbf{v}'(0)\rangle}\right) = \gamma(0) - \frac{\langle \gamma'(0), \mathbf{v}'(0)\rangle}{\langle \mathbf{v}'(0), \mathbf{v}'(0)\rangle}\mathbf{v}(0)$$

Since $l(u)$ and $j(u)$ are the same family of lines, we can write $\tau(u) = \gamma(u) + a(u)\mathbf{v}(u)$ for some C^2 scalar valued function $a(u)$. Using the definition of focusing, the linearity of the scalar product, and the fact that $\mathbf{v}(0) \perp \mathbf{v}'(0)$ (since $\mathbf{v}(0)$ is a unit vector) we find that $j(u)$ focuses about $j(0)$ at the point

$$\begin{aligned}
j\left(0, -\frac{\langle \tau'(0), \mathbf{v}'(0)\rangle}{\langle \mathbf{v}'(0), \mathbf{v}'(0)\rangle}\right) &= \tau(0) - \frac{\langle \tau'(0), v'(0)\rangle}{\langle \mathbf{v}'(0), \mathbf{v}'(0)\rangle}\mathbf{v}(0) \\
&= \gamma(0) + a(0)\mathbf{v}(0) - \frac{\langle \gamma'(0) + a'(0)\mathbf{v}(0) + a(0)\mathbf{v}'(0), v'(0)\rangle}{\langle \mathbf{v}'(0), \mathbf{v}'(0)\rangle}\mathbf{v}(0) \\
&= \gamma(0) + a(0)\mathbf{v}(0) - \frac{\langle \gamma'(0), \mathbf{v}'(0)\rangle}{\langle v'(0), \mathbf{v}'(0)\rangle}\mathbf{v}(0) - a(0)\mathbf{v}(0) \\
&= \gamma(0) - \frac{\langle \gamma'(0), \mathbf{v}'(0)\rangle}{\langle \mathbf{v}'(0), \mathbf{v}'(0)\rangle}\mathbf{v}(0) \\
&= l\left(0, \frac{\langle \gamma'(0), \mathbf{v}'(0)\rangle}{\langle \mathbf{v}'(0), \mathbf{v}'(0)\rangle}\right)
\end{aligned}$$

as desired. □

Hence, we can take $\gamma(0) = Z$. The family $l(u)$ focuses at $u = 0$ at $t_f = -\frac{\langle \gamma'(0), \mathbf{v}'(0)\rangle}{\langle \mathbf{v}'(0), \mathbf{v}'(0)\rangle}$. If $t_f = 0$, we must have $\gamma'(0) = 0$.

Remark: The point of this diversion was to show that we can choose our parameterizations γ_i so that $\gamma(0) = Z$ and $\gamma'(0) = 0$.

We now show that changing the "speed" of the parametrization of our family of lines does not affect its focusing points.

Lemma 3.4 *Let $\beta(u) \in C^2(I)$. Given a smooth family of lines $l(u,t)$, let $j(u,t) = \gamma(\beta(u)) + t\boldsymbol{v}(\beta(u))$. Then the focusing point of obtained along $l(0)$ is the same as the focusing point obtained along $j(0)$.*

Proof At $u = 0$, the family $l(u)$ focuses at $t_f = -\frac{\langle \gamma'(t), v'(t)\rangle}{\langle v'(t), v'(t)\rangle}|_{t=0}$. At $u = 0$, the family $j(u)$ focuses at

$$\begin{aligned} -\frac{\langle [\gamma(\beta(t))]', [v(\beta(t))]'\rangle}{\langle [v(\beta(t))]', [v(\beta(t))]'\rangle}|_{\beta(t)=0} &= -\frac{\|\beta'(t)\|^2 \langle \gamma'(\beta(t)), v'(\beta(t))\rangle}{\|\beta'(t)\|^2 \langle v'(\beta(t), v'(\beta(t))\rangle}|_{\beta(t)=0} \\ &= -\frac{\langle \gamma'(s), v'(s)\rangle}{\langle v'(s), v'(s)\rangle}|_{s=0} \end{aligned}$$

as desired. □

The following lemma shows that C^1 families of oriented lines remain C^1 after reflection.

Lemma 3.5 *Let $l(u)$, $u \in I$, be a C^1 family of oriented lines parameterized by $l(u,t) = \gamma(u) + t\boldsymbol{v}(u)$, where $\gamma(u) \in M$ for all $u \in I$. Assume that $\boldsymbol{v}(u)$ points inside the table if $\gamma(U)$ intersects the boundary. Let $l_1(u)$ be the oriented line obtained from $l(u)$ after one reflection. (Define this earlier–the billiard map T will be helpful). Then $l_1(u)$ is a C^1 family of lines that can be parameterized by $l_1(u,t) = \gamma_1(u) + t\boldsymbol{v}_1(u)$ where $\gamma_1(u) = \sigma(\beta(u))$ for some C^1 function $\beta : I \to S^1$ and $\boldsymbol{v}_1(u)$ points inside the table at $\sigma(\beta(u))$.*

Proof Take u to be fixed, so $\mathbf{v} := \mathbf{v}(u)$ and $a := \gamma(u)$ are fixed. Consider the sequence of maps $t \mapsto a + t\mathbf{v} \mapsto w$ coordinate of $a + t\mathbf{v}$. We call the first map f and the second g. We want to find a C^1 function $t = t(u)$ such that this function gives a value of 0. The implicit function theorem guarantees that such a function exists if $\frac{d}{dt}(g(f(t))) \neq 0$. We have $\frac{d}{dt}(g(f(t))) = \nabla g(f(t)) \cdot f'(t) = \nabla g(f(t)) \cdot \mathbf{v} = -\mathbf{w} \cdot \mathbf{v} \neq 0$, since $\mathbf{w} \perp \sigma'$ and $\mathbf{v}$ is not parallel to σ. □

Given a vertex p of a billiard path in M, we consider a family of oriented lines $l(u)$, $|u| < \delta$ with base point p, along a segment that has p as an endpoint. Each time the family $l(u)$ hits ∂M, each line in the family obeys the billiard reflection law. Let $l_k(u)$ be the resulting family after k reflections.

Definition 3.6 Let p and p' be points in M, and let τ be a billiard path from p to p'. Let $l_1, l_2, \ldots l_k$ be the oriented lines determined by the segments of τ. Then p and p' are said to be *conjugate* along τ if the following holds: there exists a family of lines $l(u)$, $|u| < \epsilon$, such that the successive reflections (from ∂M) $l_1(u), \ldots l_k(u)$ with $l(u) = l, l_1(0) = l_1, \ldots, l_k(0) = l_k$ such that $l(u)$ focuses at p at $u = 0$ and $l_k(u)$ focuses at p' at $u = 0$.

Lemma 3.7 *Suppose we have a C^1 family of oriented lines $l(u)$ parameterized by $l(u,t) = \gamma(u) + t\boldsymbol{v}(u)$. If $l(u_0, t_0)$ is not the focusing point of the family at $u = u_0$, then $Dl(u_0, t_0)$ is invertible.*

Proof We have $Dl(0,t_0) = \left(\frac{\partial l}{\partial u}, \frac{\partial l}{\partial t}\right) = (\gamma'(u_0) + t_0\mathbf{v}'(u_0), \mathbf{v}(u_0))$. Since $\mathbf{v} \perp \mathbf{v}'$, the columns of this matrix are linearly independent if and only if $0 = \mathrm{pr}_{\mathbf{v}'} Dl(0,t_0) := \frac{\langle \gamma'(u_0), \mathbf{v}'(u_0)\rangle}{\langle \mathbf{v}'(u_0), \mathbf{v}'(u_0)\rangle} + t_0$. But this happens if and only if $l(u_0,t_0)$ is the focusing point of the family $l(u)$ at $u = u_0$. Thus $Dl(u_0,t_0)$ is invertible. □

Later we will show that under certain conditions (namely (4) in 4.1) if we have a billiard path from x to y for a given table, and we make a small C^2 perturbation of the table, then there is a billiard path from x to y in the new table that is close to the one in the old table. The following lemma is needed.

Lemma 3.8 *Suppose $U \in \mathbb{R}^2$ is open, $f : U \to \mathbb{R}^2$ is C^1, and df_{x_0} is invertible at some $x_0 \in U$. Let $y = f(x_0)$. Let K be a compact set that is the closure of an open set and is such that $x \in intK \subset K \subset U$. Let $\epsilon > 0$. Then there exists $\delta > 0$ such that $dist_{C^1(K,\mathbb{R}^2)}(f,g) < \delta$ implies there exists x_1 with $dist(x_1, x_0) < \epsilon$ such that $g(x_1) = y$.*

Proof By decreasing the size of K if necessary, we may assume that the restriction of f to the interior of K is a diffeomorphism onto its image. By decreasing ϵ if necessary, we may assume that $\overline{B}_\epsilon(x_0) \subset \mathrm{int}K$. We can choose δ sufficiently small (where δ is defined in the statement of this lemma), we may assume that Dg_x is invertible for all $x \in K \supset \overline{B}_\epsilon(x_0)$. Let $C = \partial B_\epsilon(x_0)$. Since f is a diffeomorphism, we have $\mathrm{dist}(f(C), y_0) := \rho > 0$. If $\delta < \rho/2$, then

$$\mathrm{dist}_{C^0}(f,g) < \delta \implies \mathrm{dist}(g(C), y) > \rho/2 \tag{1}$$

and

$$\mathrm{dist}_{C^0}(f,g) < \delta \implies \mathrm{dist}(g(x_0), y) < \rho/2 \tag{2}$$

Let $x_1 \in \overline{B}_\epsilon(x_0)$ be such that

$$\mathrm{dist}(g(x_1), y) = \min\left\{\mathrm{dist}(g(x_1), y) : x \in \overline{B}_\epsilon(x_0)\right\} \tag{3}$$

Note that (1) and (2) imply $x \notin \partial B_\epsilon(x_0)$. But we cannot have $\mathrm{dist}(g(x_1), y) > 0$, because $g(B_\epsilon(x_0))$ is open and therefore contains an open ball around $g(x_1)$, which would contradict (3). Thus $g(x_1) = y$. □

3.3 Baire category theorem and some corollaries

We now give some definitions, state and prove the Baire category theorem, and prove some of its corollaries. Throughout this discussion, let X be a complete metric space with metric d. Up until Corollary 3.16, all sets mentioned are subsets of X.

Definition 3.9 A set S is called nowhere dense if for every open ball U, there exists an open ball $V \subset U$ such that $V \cap S = \emptyset$.

Definition 3.10 A set T is called *residual* if it is a countable union of nowhere dense sets.

We now prove the Baire category theorem.

Proposition 3.11 *X is not residual.*

Proof Suppose that $X = \bigcup_{i=1}^{\infty} S_i$, where the S_i are nowhere dense sets. Let B be a non-empty open ball; since S_1 is nowhere dense, there exists a non-empty open ball $B_1 \subset B$ such that $B_1 \cap S_1 = \emptyset$. We need the following lemma:

Lemma 3.12 *Let $r > 0$. Then $\overline{B}_r(x_0) \subseteq \{x \in X : d(x, x_0) \leq r\}$.*

Proof Let $x \in \overline{B}_r(x_0)$ and suppose that $r' := d(x, x_0) > r$; by the definition of closure, for every $\epsilon > 0$ we have $B_\epsilon(x) \cap B_r(x_0) \neq \emptyset$. Let $\epsilon = \frac{r'-r}{2}$ and let $y \in B_r(x)$; then by the triangle inequality $d(y, x_0) \geq d(x, x_0) - d(y, x) > r' - \epsilon = r' - \frac{r'-r}{2} = \frac{r+r'}{2} > r$. This means that $B_\epsilon(x) \cap B_r(x_0) = \emptyset$, a contradiction. Thus, we must have $d(x, x_0) \leq r$. □

Let $B_1 = B_{r_1}(x_1)$, where $r_1 < 1$ is sufficiently small that there exists an $\epsilon > 0$ such that $B_{r_1+\epsilon}(x_1) \cap S_1 = \emptyset$; by Lemma 3.12, $\overline{B}_{r_1}(x_1) \subset B_{r_1+\epsilon}(x_1)$, so $\overline{B}_1 \cap S_1 = \emptyset$. Now construct B_2, B_3, etc. in this way, so we get non-empty open balls $(B_i)_{i=1}^{\infty}$ such that $\overline{B}_{n+1} \subset B_n$, $B_n = B_{r_n}(x_n)$ where $r_n < 1/n$, and $B_n \cap S_n = \emptyset$ for all $n \in \{1, 2, \ldots\}$. Let $\epsilon > 0$ and take $n > 2/\epsilon$; then for all $l, m > n$ we get $B_l \subset B_n$ and $B_m \subset B_n$, so $x_l, x_m \in B_n$, i.e. $d(x_l, x_n) < 1/n < \epsilon/2$. Thus $d(x_l, x_m) < d(x_l, x_n) + d(x_m, x_m) < \epsilon$, so $(x_i)_{i=1}^{\infty}$ is a Cauchy sequence. Since X is complete, there exists $x \in X$ such that $\lim_{i\to\infty} x_i = x$. Since B_n is non-empty for all $n \in \{1, 2, \ldots\}$, for each n there exists $\epsilon > 0$ such that $r_n > \epsilon$. There exists N such that $d(x, x_m) < \epsilon/2$ for all $m > N$, and if we also take $m > 2/\epsilon$ we get $d(x, x_m) < \epsilon/2$; thus $d(x, x_n) < \epsilon$, so $x \in \bigcap_{i=1}^{\infty} B_i$. Since $x \in B_i$ and $B_i \cap S_i = \emptyset$ for all positive integers i, we have $x \notin \cup_{i=1}^{\infty} S_i$, i.e. $x \cap X = \emptyset$, which is impossible since $x \in X$. □

Corollary 3.13 *Consider the sequence $(A_i)_{i=1}^{\infty}$, where the A_i are dense open sets. Let $A = \bigcap_{i=1}^{\infty} A_i$. Then A is dense.*

Proof Suppose that A is not dense in X, i.e. there exists a non-empty open ball B such that $B \cap A = \emptyset$. Then $X = (B \cap A)^c = B^c \cup A^c = B^c \cup A_1^c \cup A_2^c \cup \cdots$, so we must have $B \subset \bigcup_{i=1}^{\infty} A_i^c$. The complement of a nowhere dense set is dense, and a subset of a countable union of nowhere dense sets is itself a countable union of nowhere dense sets. Thus, we can write B as a countable union of nowhere dense sets.

We now prove two lemmas. Let $B' = B_r(x_0) \subset X$ be an open ball.

Lemma 3.14 *$(\overline{B'}, d)$ is a complete metric space.*

Proof Let $(b_i)_{i=1}^{\infty}$ be a Cauchy sequence in $\overline{B'}$. Since $b_i \in X$ for all positive integers i, and X is complete, we find that there exists $x \in X$ such that $\lim_{i\to\infty} b_i = x$. Suppose that $x \notin \overline{B'}$. Since $\overline{B'}$ is closed, there exists $\epsilon > 0$

such that $B_\epsilon(x) \subset \left(\overline{B'}\right)^c$. Since $(b_i)_{i=1}^\infty$, there exists N sufficiently large that $b_i \in B_\epsilon(x)$. But $b_i \in \overline{B'}$, so we have a contradiction. Thus $x \in \overline{B'}$, so $\overline{B'}$ is complete. □

Lemma 3.15 *B' is not residual.*

Proof By Lemma 3.12, we have $\overline{B}_{r/2}(x_0) \subset B_r(x_0)$. By Lemma 3.14, we find that $\overline{B}_{r/2}(x_0)$ is not a countable union of nowhere dense sets, so neither is $B_r(x_0) = B'$. □

B was assumed to be an open ball, and we have found that if A is not dense then we can express B as a countable union of nowhere dense sets. This contradicts Lemma 3.15. □

Let (Y, d') be a metric space that is not necessarily complete. For $n \in \{1, 2, \ldots\}$, let $Y_n \subset Y$ be a complete metric space (with metric d') that is the closure of an open set, and such that $Y = \bigcup_{n=1}^\infty Y_n$. Suppose that for each $i \in \{1, 2, \ldots\}$, A_i is a dense open subset of Y.

Corollary 3.16 *$\bigcap_{i=1}^\infty A_i$ is dense in Y.*

Proof Since A_i is dense in Y, it is dense in each Y_n. By the same argument as in the proof of Corollary 3.13, we find that $Y_n \cap A_1 \cap A_2 \cap \cdots$ is dense in Y_n for all positive integers n. Hence $\bigcap_{i=1}^\infty A_i$ is dense in Y. □

In this paper, we are interested in metric spaces of the form $C^2(U, V)$, where U is a compact set and V is either $\mathbb{R}$ or $\mathbb{R}^2$. The metric space $C^2(U, V)$ consists of functions with continuous second derivatives, and the metric is defined as $d(f, g) = \sup_{x \in U}[|f(x) - g(x)| + |f'(x) - g'(x)| + |f''(x) - g''(x)|]$.

4 The Main Result

Our goal is to prove the following.

Theorem 4.1 *Given any positive integer n and any $\epsilon > 0$, there exists a C^2 perturbation of σ, of size less than ϵ in the C^2 topology, such that in the perturbed billiard table, there are n billiard paths from x to y with no triple intersections except at x and y.*

To prove Theorem 4.1, we proceed by induction. It is convenient to assume some additional conditions as part of the inductive hypothesis. The complete inductive hypothesis consists of the following conditions:

1. There is no periodic path that uses only vertices from these n paths.
2. No two paths share a vertex.
3. These paths have no triple intersections except at x and y.
4. The points x and y are not conjugate along any of these paths. (See Definition 3.6.)

The base case (one path from x to y) is trivial. Suppose that $n > 1$ is an integer and there are n paths satisfying $(1) - (4)$. We want to show that there exists an $(n+1)$st path from x to y that is distinct from the first n paths and satisfies conditions $(1) - (4)$.

5 Proof of Theorem 4.1

We proceed in steps toward a proof of Theorem 4.1, as outlined in Section 2.

Let $\mathcal{P} = \{p_1, p_2, \ldots p_k\}$ be the vertices of the first n paths. Let N be an integer greater than $2p_k^2$; clearly N is greater than the number of pairs of vertices in $\mathcal{P}$. We take our $(n+1)$st path to be the maximum length path having N vertices, which is a billiard path by Lemma 3.1. Since there are no periodic paths using vertices in $\mathcal{P}$, our $(n+1)$st path must have at least one vertex not in $\mathcal{P}$; if it did not, then by the pigeonhole principle the $(n+1)$st path would contain some pair of vertices twice, forcing it to be periodic and contradicting condition (1).

Consider the portion γ of the curve σ in a small neighborhood of the vertex p. Let the family of oriented lines $l(u)$ be reflected off γ to give the family $l_1(u)$. Parametrize these families as $l(u,t) = \gamma(u) + t\mathbf{v}(u)$ and $l_1(u,t) = \gamma_1(u) + t\mathbf{v}_1(u)$, where $|u| < \delta$. Let $\alpha(u) \in (0, \pi)$ be the angle made by $\mathbf{v}_1(u)$ with respect to the tangent line at $\gamma(u)$, and let $\kappa(u)$ be the signed curvature of γ at u (the curvature has a sign after we choose the direction of the normal vector $\mathbf{N}(u)$). Lemma 1 in [9] states that if $t = f(s)$ is the local envelope of $l(u)$ and $t = f_1(u)$ is the local envelope of the reflected family $l_1(u)$, then $-1/f(u) + 1/f_1(u) = 2\kappa(u)/\sin\alpha$. Thus if p and p' are conjugate vertices along a segment of some billiard path, then by changing the curvature of σ at p, we are changing $\kappa(u)$ while leaving $f(u)$ fixed, which ensures that $f_1(u)$ changes. Even if many reflections are required to get from p to p' on the billiard path, it is evident from the equation in [9] that by changing the curvature of σ at p, we ensure that p and p' are no longer conjugate.

Now we state a proposition that allows us to avoid unwanted focusing of families of oriented lines along some billiard path. Suppose that p is a vertex of a billiard path such that the when the family of rays around this path reflects off the boundary in a small neighborhood of p, it focuses at some point in the table. See Figures 2 and 3.

Proposition 5.1 *Given $\epsilon > 0$, we can perturb σ in an ϵ-neighborhood of p in such a way that the position of p and the tangent line to σ at p are unchanged, and the new table σ_1 is still C^2 and strictly convex, the curvature of σ_1 at p is different from the curvature of σ at p.*

Now we state and prove the calculus version of Proposition 5.1.

Lemma 5.2 *Let $f : [-\nu, \nu] \to \mathbb{R}$ (where $\nu > 0$) be a C^2 strictly convex function describing part of the boundary of the strictly convex C^2 billiard table $\sigma : S^1 \to \mathbb{R}$, with Cartesian coordinates imposed so that $f(0) = f'(0) = 0$. Let $b \in (0, \nu)$.*

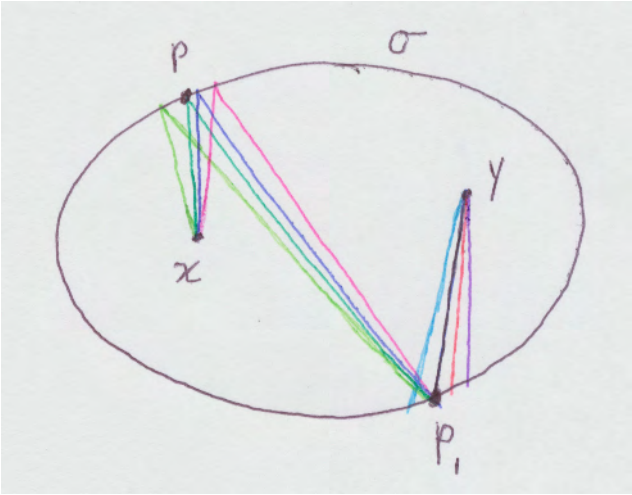


Figure 2: A problem with focusing at P

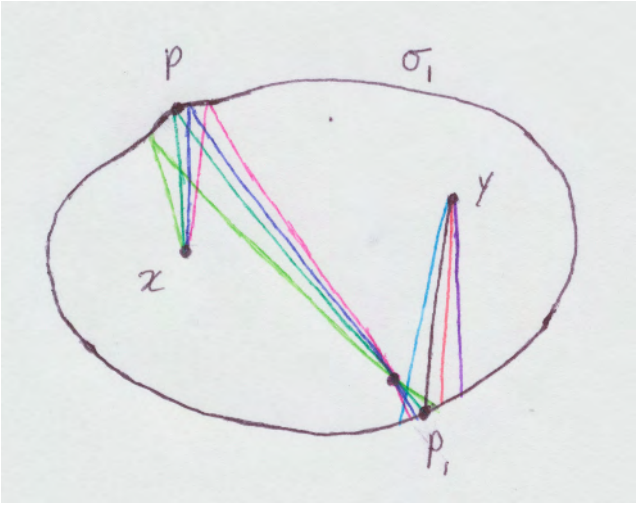


Figure 3: Rays no longer focus at P

There exists a smooth strictly convex C^2 function $g : \mathbb{R} \to \mathbb{R}$ such that $f(x) = g(x)$ for all $x \in [-\nu, -b] \cup [b, \nu]$, $g(0) = g'(0) = 0$, and $g''(0) \neq f''(0)$.

Proof We construct g by adding a C^∞ strictly convex "bump function" $\psi : \mathbb{R} \to \mathbb{R}$ that satisfies $|\psi(x)| < a$ for all $x \in \mathbb{R}$. We write

$$g(x) = f(-c) + \int_{-c}^{t} \left[f'(-c) + \int_{-c}^{s} (f''(s) + \psi(s)) ds \right] dt$$

Then $g(0) = \int_{-c}^{0} \int_{-c}^{t} \psi(x) dx dt$ and $g'(0) = \int_{-c}^{0} \psi(x) dx$. Let $\Psi : \mathbb{R} \to \mathbb{R}$ be such that $\Psi'(x) = \psi(x)$ for all $x \in \mathbb{R}$. By Fubini's theorem and integration by parts we get

$$\begin{aligned} g(0) &= \int_{-c}^{0} \int_{-c}^{t} \psi(x) dx dt = \int_{-c}^{0} \int_{x}^{0} \psi(x) dt dx = \int_{-c}^{0} (-x\psi(x)) dx \\ &= (-x\Psi(x))|_{-c}^{0} + \int_{-c}^{0} \Psi(x) dx \\ &= c\Psi(-c) + \int_{-c}^{0} \Psi(x) dx \end{aligned}$$

And thus we want to have $\Psi(-c) = 0$, $\int_{-2c}^{0} \Psi(x) = 0$, and $\Psi(0) = 0$ (since we need $g'(0) = 0$, i.e. $\Psi(0) - \Psi(-c) = 0$). We also need $|\psi(x)| < a$, which we ensure by taking $|\Psi'(x)| < a$ for all $x \in \mathbb{R}$, i.e. Ψ should not increase too quickly. We can construct Ψ to have the desired properties by defining it piecewise in terms of several bump functions, and then we take $\psi := \Psi'$. □

Using Lemma 1 in [9] (sometimes known as the "mirror equation") and Proposition 5.1, we change the table slightly a finite number of times to ensure that no two vertices of the $(n+1)$st path are conjugate along that path, and x and y are not conjugate along that path.

We now state the proposition that is the "workhorse" in the proof of 4.1. Consider a strictly convex C^2 billiard table with boundary $\sigma : S^1 \to \mathbb{R}^2$. Consider a billiard path APB, where P is a vertex, A and B are interior points close to P, and the path is not perpendicular to the table. See Figure 4.

Proposition 5.3 *If P' is sufficiently close to P and the line l is nearly parallel to the tangent line to σ at P, then we can perturb σ in a small neighborhood of P to get a new boundary σ_1 in such a way that the tangent line to σ_1 at P' is l, and σ_1 is C^2 and strictly convex.*

Now we state and prove the calculus version of Proposition 5.3 and some corollaries. Let the billiard path APB be on a section of the table described by the C^2 strictly convex function $f : [-\nu, \nu] \to \mathbb{R}$ (where $\nu > 0$) with coordinate system taken so that $f(0) = f'(0) = 0$. Let $b \in (0, \nu)$.

Lemma 5.4 *Given $\epsilon > 0$, there exists δ (depending on ϵ, b, and ν) such that if $|P' - P| < \delta$ and $|m| < \delta$, then there exists $g \in C^2[-\nu, \nu]$ such that the graph of g passes through P' with slope m, $f(x) = g(x)$ for all $x \in [-\nu, -b] \cup [b, \nu]$, and $\|f - g\|_2 < \epsilon$ (so g is strictly convex).*

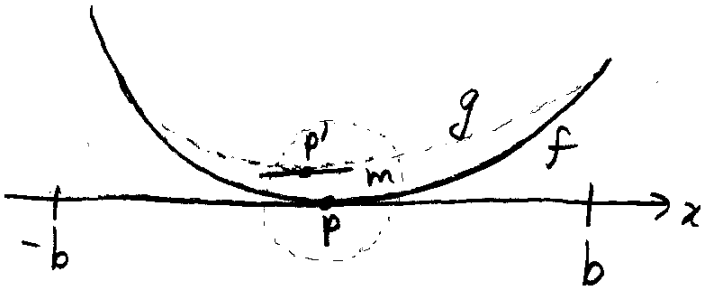


Figure 4: Lemma 5.4

Proof We write $g(x) = f(-b) + \int_{-b}^{x}[f'(-b) + \int_{-b}^{t}(f''(s) + \psi(s))ds]dt$, where $\psi : [-\nu, \nu] \to \mathbb{R}$ is a C^∞ function such that $|\psi(x)| < a$ for all $x \in [-b, b]$, $\psi(x) = 0$ for all $x \in [-\nu, -b] \cup [b, \nu]$, and $\int_{-b}^{b} \psi(x)dx = 0$. Let $P' = (c, d)$. We wish to have $g(c) = d$ and $g'(c) = m$. Using Fubini's theorem and integration by parts, we get

$$\begin{aligned} g(c) &= f(c) + \int_{-b}^{c}\int_{-b}^{t} \psi(s)dsdt = f(c) + \int_{-b}^{c}\int_{s}^{c} \psi(s)dtds \\ &= f(c) + \int_{-b}^{c} (c\psi(s) - s\psi(s))ds = f(c) + c(\Psi(c) - \Psi(-b)) - \int_{-b}^{c} s\psi(s)ds \\ &= f(c) + c\Psi(c) - \int_{-b}^{c} s\psi(s)ds \\ &= f(c) + c\Psi(c) - (s\Psi(s)|_{-b}^{c} + \int_{-b}^{c} \Psi(s)ds \\ &= f(c) + \int_{-b}^{c} \Psi(x)dx \end{aligned}$$

and

$$g'(c) = f'(c) + \int_{-b}^{c} \psi(s)ds = f'(c) + \Psi(c)$$

We require $\Psi'(x) = \psi(x)$ for all $x \in [-\nu, \nu]$, $\Psi(-b) = \Psi(b) = 0$, and $|\Psi'(x)| < a$ for all $x \in [-\nu, \nu]$. Thus we need

$$|m - f'(c)| = |\Psi(c)| = \left|\int_{-b}^{c}\int_{-b}^{x} \Psi'(t)dt\right| < a(b + c)$$

and

$$\begin{aligned} |d - f(c)| &= \left|\int_{-b}^{c} \Psi(x)dx\right| = \left|\int_{-b}^{c}\int_{-b}^{x} \Psi'(t)dtdx\right| \\ &< \left|\int_{-b}^{c} a(b + x)dx\right| = |ab(b + c) + \frac{1}{2}a(c^2 - b^2)| \end{aligned}$$

so we need $\frac{|m-f'(c)|}{b+c} < a$ and $\frac{2|d-f(c)|}{(b+c)^2} < a$. As $m \to 0$, $c \to 0$, and $d \to 0$, we have $f(c) \to 0$, and it is clear that the above quotients approach 0. Thus for sufficiently small values of m, c, and d, they are bounded above by a, as desired. □

Consider a billiard path PAP_1, where A is a vertex, P and P_1 are interior points, and the path is not perpendicular to ∂M.

Corollary 5.5 *For any A' on the line AP_1 sufficiently close to but not the same as A, we can perturb σ in a small neighborhood of A to get σ_1 in such a way that σ_1 is a strictly convex C^2 curve and $PA'P_1$ is a billiard path.*

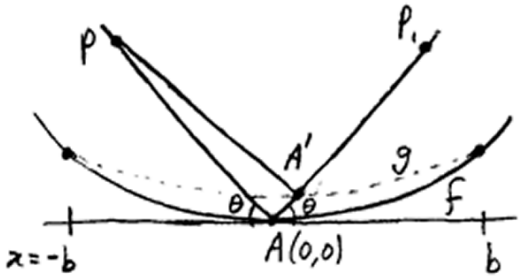


Figure 5: Eliminating a periodic path

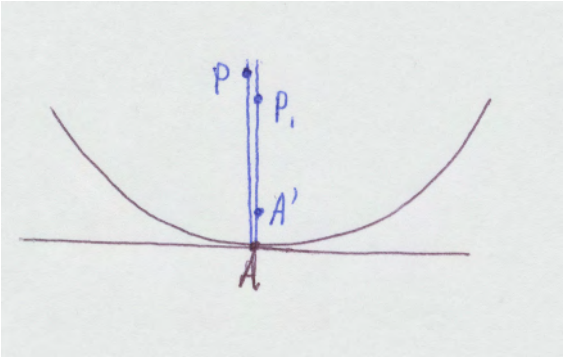


Figure 6: A perpendicular path creates a problem

Now we state and prove the calculus version of Corollary 5.5. Let σ, f, and b be as in Lemma 5.4. Consider the billiard path PAP_1 shown in Figure 5, with angles of incidence and reflection $\theta \in (0, \pi/2)$. If $\theta = \pi/2$, then since A' is on AP in this case, we cannot change the table to simultaneously alter θ and keep the billiard path going through P and P'. See Figure 6.

Corollary 5.6 *For any A' on the line AP_1 that is sufficiently close, but not equal, to A, there exists a strictly convex C^2 function $g : [-\nu, \nu] \to \mathbb{R}$ such that the angles made by PA' and $A'P_1$ with respect to the tangent line to g at A' are equal (i.e. $PA'P_1$ is a billiard path), and $g(x) = f(x)$ for all $x \in [-\nu, -b] \cup [b, \nu]$.*

Proof It is clear that the required slope of the tangent line to g at A' approaches 0 as $|A - A'| \to 0$. Now we can apply Proposition 5.3 to get the desired result. □

Now we state another corollary. Let σ be as before, and let P_1APBP_2 be a billiard path, where A and B are vertices, P_1, P, and P_2 are interior points, and none of the segments of this path are perpendicular to the boundary.

Corollary 5.7 *For any A' sufficiently close (but not equal) to A, we can perturb σ in small neighborhoods of A and B to get σ_1 in such a way that σ_1 is a strictly convex C^2 curve, and $P_1A'B'P_2$ is a billiard path, where B' is the point where the line through A' parallel to AP intersects σ_1.*

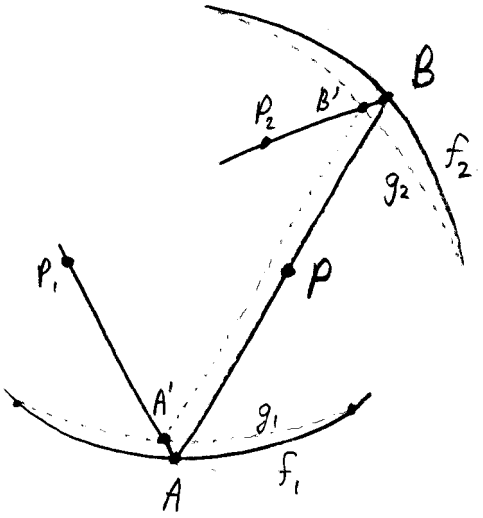


Figure 7: Eliminating a triple intersection through P

Now we state and prove the calculus version of Corollary 5.7. Let σ and b be as in 5.4, and suppose that $f_1 : [-\nu, \nu] \to \mathbb{R}$ and $f_2 : [-\nu, \nu] \to \mathbb{R}$ (where $\nu > 0$) are functions that are locally graphs of σ in neighborhoods of A and B, respectively. We introduce two Cartesian coordinate systems so that in the first, $f_1(0) = f_1'(0) = 0$, and in the second, $f_2(0) = f_2'(0) = 0$. Consider the billiard path P_1APBP_2 shown in Figure 7, and suppose that AB is not perpendicular to the tangent line to the boundary at A or B.

Corollary 5.8 *Then for any A' on P_1A sufficiently close but not identical to A, and for any B' on P_2B sufficiently close but not identical to B, with $A'B'$ parallel to AB, there exist strictly convex C^2 functions $g_1 : [-\nu, \nu] \to \mathbb{R}$ and $g_2 : [-\nu, \nu] \to \mathbb{R}$ such that $P_1A'B'P_2$ is a billiard path.*

Proof It is clear from Figure 7 that the required slope of the tangent line to g_1 at A' approaches 0 as $|A - A'| \to 0$, and similarly the required slope of the tangent line to B' approaches 0 as $|B - B'| \to 0$. The corollary now follows from Proposition 5.3. □

Consider families of rays from x to y along the $(n+1)$st path. Since there are uncountably many such rays and x and y are not conjugate to any of the (finite number of) vertices, we can find uncountably many paths that hit none of the old vertices, which have no segments perpendicular to the boundary, and do not pass through x of y except at the beginning and end (i.e. x and y are not interior points of any of these paths). However, we still need to show that there exist pairs of rays (one ray from x and the other from y) such that one of their angle bisectors has slope arbitrarily close to the slope of the tangent line at the new vertex, and the intersection point of these rays is arbitrarily close to the new vertex.

Definition 5.9 Let l_1 and l_2 be oriented lines with direction vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, respectively, where $\mathbf{v}_1$ and $\mathbf{v}_2$ are unit vectors and $\mathbf{v}_1 \neq \mathbf{v}_2$. Then the angle bisector of l_1 and l_2 is the oriented line with direction vector $\mathbf{v}_1 + \mathbf{v}_2$.

Lemma 5.10 *Let l be an oriented line with direction vector $\boldsymbol{v}$, and let Z_1, Z_2, and P be distinct points on l. Suppose $l_1(u)$ and $l_2(u)$, where $-\delta < u < \delta$, are smooth families of oriented lines with direction vectors $\boldsymbol{v}_i(u)$ such that $l = l_1(0) = l_2(0)$ for $i = 1, 2$. Assume that at $u = 0$, l_i focuses at Z_i for $i \in \{1, 2\}$. Then for every $\epsilon > 0$, there exist oriented lines $L_i = L_i(\epsilon) \in \{l_i(u), -\delta < u < \delta\}$ ($i \in \{1, 2\}$) such that L_1 and L_2 intersect at a unique point P', and P' satisfies $|P' - P| < \epsilon$. Moreover, the direction vectors $\boldsymbol{V}_i$ of L_i satisfy $|\boldsymbol{V}_i - \boldsymbol{v}| < \epsilon$.*

Proof Without loss of generality, we may assume that $\mathbf{v} = (1, 0)$, $P = (0, 0)$, and $Z_i = (z_i, 0)$. We parameterize l_i by $l_i(u) = \gamma_i(u) + t\mathbf{v}_i(u)$. According to the definition before Lemma 3.3, if $f(u) = -\langle \gamma_i', v_i' \rangle / \langle v_i', v_i' \rangle$ (where $\mathbf{v}_i' \neq \mathbf{0}$), then $l_i(u, f(u))$ is the point on $l_i(u)$ where the family $l_i(u)$ is focused in linear approximation at $u = 0$. We will show that for every $\epsilon > 0$ there exist u_1, u_2 such that $l_1(u_1) = l_2(u_2) = (0, d)$ where $|d| < \epsilon$, and the slope m of $\mathbf{v}$ satisfies $|m| < \epsilon$.

By re-parameterizing (and possibly reversing the direction of the parameterization), we may assume $\mathbf{v}_i(u)$ has slope u. Thus, we write $\gamma_i(u) = Z_i + o(u) = (z, 0) + (o(u), o(u)) = (z + o(u), o(u))$. Here we define $h(u) = o(u)$ if $h : \mathbb{R} \to \mathbb{R}$ is a function such that $h(u)/u \to 0$ as $h \to 0$.

We write $\gamma_i(t)$ in coordinates as $(\gamma_i^1(t), \gamma_i^2(t))$. The line $\gamma_i(u) + t\mathbf{v}_i(u)$ hits the y-axis at $y = \gamma_i^2(u) - u\gamma_i^1(u) = u(-z + o(u))$. As $u \to 0$, this expression lies between $-2zu$ and $-\frac{z}{2}u$. Depending on the sign of u, it is on the positive or negative side of the y-axis. Since the y-intercept is a continuous function of u,

by the intermediate value theorem there exists u for which the line $\gamma_i(u)+t\mathbf{v}_i(u)$ hits the y-axis for all $y \in (\frac{z}{2}u, -\frac{z}{2}u)$.

We have found that the y-intercept of the line depends on u, the line's slope, through the function $u \mapsto u(-z + o(u)) = -zu + o(u)$. Thus the derivative of the y-intercept as a function of u is $-z$, so the derivative of the inverse function is $-\frac{1}{z}$. Call this function m; then m maps y to slope. Since $m'(t) = -1/z$, we have $m(y) = (-\frac{1}{z})y + o(y)$.

From the argument earlier, we know that there exist u_1 and u_2 such that $\gamma_1(u_1)$ and $\gamma_2(u_2)$ both intersect the y-axis at $(0, y_0)$. Let the slopes of these lines be m_1 and m_2 respectively; then by the formula above we have $m_1 = -\frac{1}{z_1} + o(u)$ and $m_2 = -\frac{1}{z_2}y_0 + o(y_0)$. For sufficiently small values of y_0, these slopes are not equal. Hence $\lim_{u\to 0} m(u) = 0$. Observe that if $\mathbf{V}$ has slope less than ϵ in absolute value, then $|\mathbf{V} - (1, 0)| < \epsilon$, as desired. □

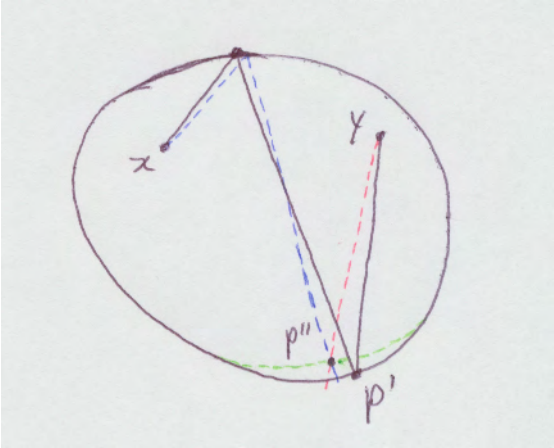


Figure 8: Matching up paths

Now we can use Proposition 5.3 to "match up" two of the paths we showed exist in Lemma 5.10 by changing the table in a neighborhood of the new vertex. This gives us a new $(n+1)$st billiard path with no vertices in $\mathcal{P}$. See Figure 8.

Now we use Corollary 5.5 to eliminate periodic paths that use vertices from any of the $(n+1)$ paths. We then use Corollary 5.7 to eliminate triple intersections (except those at x and y) that have arisen from constructing the $(n+1)$st path. Now we have constructed $(n+1)$ billiard paths from x to y that satisfy conditions $(1) - (4)$, and have made only small changes to the boundary of the table. This completes the proof of 4.1.

Suppose that ξ and $\tilde{\xi}$ are polygonal paths from x to y with the same number of vertices. Let the vertices of ξ be $p_1, \ldots p_k$ and the vertices of $\tilde{\xi}$ be $\tilde{p_1}, \ldots \tilde{p_k}$. We define $d(\xi, \tilde{\xi}) = \max\{d(p_i, \tilde{p}_i) : i \in \{1, \ldots k\}\}$.

Lemma 5.11 *Consider a strictly convex C^2 curve $\tau : S^1 \to \mathbb{R}^2$ and points x and y in the interior of the region bounded by τ. Suppose there exist n billiard*

paths $\xi_1, \dots \xi_n$ for τ from x to y with no triple intersections except at x and y and no common vertices, and x and y are not conjugate along any of these paths. Then there exists an open neighborhood $\mathcal{N}$ of τ in $C^2(S^1, \mathbb{R}^2)$ such that for every $\alpha \in \mathcal{N}$, x and y are still in the interior of the region bounded by α, and there exist n billiard paths $\tilde{\xi}_1, \dots \tilde{\xi}_n$ for α from x to y with no triple intersections except at x and y and no common vertices.

Proof Let $\xi_1, \dots, \xi_n$ be billiard paths for τ as above, and let $\widehat{\xi}_1, \dots \widehat{\xi}_n$ be polygonal paths from x to y such that each $\widehat{\xi}_i$ has the same number of vertices as ξ_i, and $d(\xi_i, \widehat{\xi}_i) < \epsilon$ for all $i \in \{1, \dots n\}$. We claim that if $\epsilon > 0$ is sufficiently small, then $\widehat{\xi}_1, \dots \widehat{\xi}_n$ also have no triple intersections (except at x and y) and no common vertices. Assume that i, j, k are distinct. For $\epsilon > 0$ sufficiently small, the sets $\xi_i \cap \xi_j$ and $\widehat{\xi}_i \cap \widehat{\xi}_j$ are close, and likewise $\xi_j \cap \xi_k$ and $\widehat{\xi}_j \cap \widehat{\xi}_k$ are close. Since $\xi_i \cap \xi_j \cap \xi_k = \emptyset$, it follows that for ϵ sufficiently small we have $\widehat{\xi}_i \cap \widehat{\xi}_j \cap \widehat{\xi}_k = \emptyset$. Similarly, if $i \neq j$ and $\epsilon > 0$ is sufficiently small, then $\widehat{\xi}_i$ and $\widehat{\xi}_j$ have no common vertices.

Next we show that given $\epsilon > 0$, there exists an open neighborhood $\mathcal{N}$ of τ sufficiently small that for all $\alpha \in \mathcal{N}$, there exist billiard paths $\tilde{\xi}_1, \dots \tilde{\xi}_n$ for α from x to y such that $d(\tilde{\xi}_i, \xi_i) < \epsilon$. Let $\xi = \xi_i$ for some $i \in \{1, \dots, n\}$. Let $l(u)$, $|u| < \eta$, be a family of oriented lines parameterized by $l(u,t) = \gamma(u) + t\mathbf{v}(u)$, where $\gamma(u) \equiv x$ and the initial segment of ξ is contained in $l(0)$. Suppose ξ makes k reflections on the table bounded by τ and goes from x to y. Let $l_1(u), \dots l_k(u)$, $|u| < \eta$, be the families of oriented lines obtained by k reflections of $l(u)$ on the table bounded by τ. Now consider the families $\tilde{l}_1(u), \dots, \tilde{l}_k(u)$, $|u| < \eta$, obtained by reflecting $l(u)$ on the table bounded by α, where $\alpha : S^1 \to \mathbb{R}^2$ is C^2 close to τ.

Let $f(u,t) = l_k(u,t)$ and $g(u,t) = \tilde{l}_k(u,t)$. The final segment of ξ is contained in $l_k(0)$, and there exists t_0 such that $f(0, t_0) = y$. We can make g as C^1 close to f as desired by taking α sufficiently C^2 close to τ (a bit of additional argument is required here). By condition (4) in Theorem 4.1, y is not the focusing point for $l_k(u)$ at $u = 0$. Then by Lemma 3.7, $Df(0, t_0)$ is invertible. It follows from Lemma 3.8 that there exists (u_1, t_1) close to $(0, t_0)$ such that $g(u_1, t_1) = y$. The point (u_1, t_1) can be taken as close to $(0, t_0)$ as desired by taking α sufficiently C^2 close to τ. We let $\tilde{\xi}$ be the billiard path determined by $\tilde{l}(u_1), \tilde{l}_1(u_1), \dots \tilde{l}_k(u_1)$ that goes from x to y for the table bounded by α. By choosing $\alpha \in \mathcal{N}_i$, where $\mathcal{N}_i$ is a sufficiently small neighborhood of τ in the C^2 topology, we can make $\text{dist}(\xi, \tilde{\xi}) < \epsilon$. We do this for each $i \in \{1, \dots, n\}$, and take $\mathcal{N} = \bigcap_{i=1}^n \mathcal{N}_i$, which is the desired neighborhood of τ in $C^2(S^1, \mathbb{R}^2)$. □

We have shown that for every ϵ-neighborhood of our given table's boundary σ, there is a strictly convex C^2 closed curve τ and a $\delta_n > 0$ such that for any strictly convex C^2 closed curve α in the δ-neighborhood of τ, (2) and (3) hold for n and the table with boundary α. This shows that for fixed n, there is a C^2 dense open set G_n of boundary curves α such that (2) and (3) hold.

Let $\mathcal{C}(x, y) \subset C^2(S^1)$ consist of strictly convex curves with x and y in the interior. The set $\mathcal{C}(x, y)$ is not quite complete (due to strict convexity and the fact that x and y are in the interior). We can, however, write $\mathcal{C}(x, y)$ as a union

of complete metric spaces: $\mathcal{C}(x,y) = \bigcup_{i=1}^{\infty} \mathcal{C}_i$, where

$$\mathcal{C}_i = \left\{ \sigma \in \mathcal{C}(x,y) : \inf_{s \in S^1} \sigma''(s) \geq \frac{1}{i},\ \mathrm{dist}(x, \sigma(S^1)) \geq \frac{1}{i},\ \mathrm{dist}(y, \sigma(S^1)) \geq \frac{1}{i} \right\}$$

By Corollary 3.16, we see that $G := \bigcap_{n=1}^{\infty} G_n$ is a dense G_δ set of strictly convex C^2 boundary curves with x and y in the interior for which there exist n billiard paths from x to y that satisfy (2) and (3). This is the "generic" set of billiard tables that we wanted to find.

References

[1] M. Gerber and W.-K. Ku, *A dense G-delta set of Riemannian metrics without the finite blocking property*, Math. Res. Let. **18** (2011), no. 3, 389-404.

[2] E. Gutkin, *Billiard dynamics: an updated survey with the emphasis on open problems*, Chaos **22**, 026116 (2012).

[3] S. Kerckhoff, H. Masur, and J. Smillie, *Ergodicity of billiard flows and quadratic differentials*, Ann. Math. **124** (1986), no. 2, 293-311.

[4] T. Monteil, A Counter-example to the theorem of Hiemer and Snurnikov, J. Stat. Phys. **114** (2004), 1619-1623.

[5] S. Tabachnikov, *Birkhoff billiards are insecure*, Discrete Contin. Dyn. Syst. **23** (2009), no. 3, 1035-1040.

[6] S. Tabachnikov, *Geometry and Billiards*, AMS, 2005.

[7] Ya. Vorobets, *On the measure of the set of periodic points of a billiard*, Math. Notes **55**, 455-460 (1994).

[8] F. Warner, *Foundations of Differentiable Manifolds and Lie Groups*, Foresman Scott, New York, 1971, pp. 10-11.

[9] M. Wojtkowski, *Principles for the design of billiards with nonvanishing Lyapunov exponents*, Comm. Math. Phys. **105** (1986), no. 3, 391-414.

Generalized cyclotomy for finding supplementary difference sets in the $2p$ case

Yancy Liao

Abstract

Supplementary difference sets in the ring of residue classes modulo n have a special relationship with compatible binary sequences. Finding suppplementary difference sets of certain parameters can give us compatible binary sequences that have application in the construction of maximal determinant binary matrices. We are interested in finding these pairs of supplementary difference sets using the theory of cyclotomy. Regular cyclotomy has long been used to find supplementary difference sets either by computer search or by explicit construction, but only applies when n is prime. We considered a generalized version of cyclotomy in the $n = 2p$ (twice a prime) case. This paper gives an overview of the relationship between supplementary difference sets and compatible sequences, and their respective equivalence operations. Then we analyze the structure of the generalized cosets with regard the $n = 2p$ case. This analysis reveals how the regular and generalized cosets react to the various equivalence operations, and explicit formulas are given so that we can generate all equivalent SDS pairs starting from an initial SDS pair using the formulas alone. If we are given a collection of SDS pairs, these formulas allow us to classify them up to equivalence. Finally, we analyze the relationships between the parameters of an SDS in the $n = 2p$ case and derive 1). a new constraint on the parameters of an SDS pair, and 2). a new constraint on the coset combinations of an SDS pair. These constraints should improve the speed and efficiency of computer searches, which has been a limiting factor in the discovery of supplementary difference sets of large sizes.

1 Difference sets

Definition 1.1 [Gy1] Let $S_1, S_2, \ldots, S_e$ be subsets of a finite abelian group $(G, +)$ with $|G| = n$ and $|S_i| = k_i$. Let $\Delta S_i = \{a-b : a, b \in S_i\,,\, a \neq b\}$ be the multiset of differences for S_i. If there exists λ such that, for each $x \in G \setminus \{0\}$, x occurs in $(\bigcup_{i=1}^{e} \Delta S_i)$ exactly λ times, then $S_1, S_2, \ldots, S_e$ are called **supplementary difference sets** (SDS) with parameters $e - \{n\,;\, k_1, k_2, \ldots, k_e\,;\, \lambda\}$.

If $e = 1$, then S_1 is called an **ordinary difference set**. Letting $S_1 = G$, we can always get a trivial ordinary difference set with parameters $1-\{n\,;\, n\,;\, n-1\}$. On the other hand, we can let S_1 be a singleton set and then we get a trivial

ordinary difference set with $\lambda = 0$. Frequently we have that $k_1 = k_2 = \ldots = k_e = k$, and if so then the parameters can be abbreviated as $e - \{n\,;\, k\,;\, \lambda\}$.

Proposition 1.2 *Suppose* $e - \{n\,;\, k_1, k_2, \ldots, k_e\,;\, \lambda\}$ *are the parameters of supplementary difference sets* $S_1, S_2, \ldots, S_e$ *for some group* G*. Then*

$$\lambda(n-1) = \sum_{i=1}^{e} k_i(k_i - 1) \tag{1}$$

Proof On the left-hand side, each of the $n-1$ non-zero elements occurs λ times by assumption. On the right-hand side, if $|S_i| = k_i$, then there are $k_i(k_i - 1)$ ordered pairs from which to form differences. Then sum over $i = 1, 2, \ldots, e$. □

Corollary 1.3 *If* $|S_i| = k$ *for all* i*, then* (1) *becomes* $\lambda(n-1) = ek(k-1)$.

1.1 Equivalences

We can perform some operations on SDS that preserve the fact that they are SDS, while leaving their parameters fixed.

Definition 1.4 For subset S and $l \in G$, **addition** is $S + l = \{x + l : x \in S\}$.

Lemma 1.5 $\Delta(S_i + l) = \Delta S_i$.

Proof Given $x, y \in S_i$, $\Delta S_i \ni x - y = (x + l) - (y + l) \in \Delta(S_i + l)$ □

Theorem 1.6 *Given supplementary difference sets* $S_1, S_2, \ldots, S_e$*, adding an element to any of the sets will preserve the fact that they are SDS.*

Proof By Lemma 1.5, addition does not change any ΔS_i. □

Addition clearly preserves the parameters of SDS, so adding an element l to any of the sets is considered to be an **additive equivalence** of SDS.

There are additional equivalences if G is a ring (has multiplicative structure).

Definition 1.7 For subset S and $d \in G$, **multiplication** is $dS = \{dx : x \in S\}$.

Theorem 1.8 *Assume* $S_1, S_2, \ldots, S_e$ *are SDS. If* G *is a ring and* $d \in G$ *is an invertible element, then* $dS_1, dS_2 \ldots, dS_e$ *are SDS.*

Proof Fix non-zero $x \in G$. For $y, z \in S_i$ such that $y - z = x$, $dy - dz = dx$. So if x occurs λ_i times in ΔS_i, then dx occurs at least λ_i times in ΔdS_i. Conversely, since $\exists d^{-1}$, if $dx = dx'$ for some non-zero x', then $x = x'$. Therefore dx occurs no more than λ_i times in ΔdS_i. Iterating over i, if λ is the repetition number of x in $(\bigcup_{i=1}^{e} \Delta S_i)$, then λ is also the repetition number of x in $(\bigcup_{i=1}^{e} \Delta dS_i)$. □

Multiplication by invertible d clearly preserves the parameters of SDS, so multiplying d to every set is considered a **multiplicative equivalence**.

Multiplication by **-1** is special because we can multiply any number of the sets (not necessarily all of them) by **-1** while preserving SDS (and parameters).

Definition 1.9 Given supplementary difference sets $S_1, S_2, \ldots, S_e$, multiplying some sub-collection by $-\mathbf{1}$ is called **half-multiplication**.

Theorem 1.10 *Half-multiplication preserves SDS and is an equivalence.*

Proof If $x, y \in S_i$, then $-x, -y \in -S_i$. So if $x-y, y-x \in \Delta S_i$, then $-x-(-y) = y - x \in \Delta(-S_i)$ and $-y-(-x) = x - y \in \Delta(-S_i)$. Hence $\Delta S_i = \Delta(-S_i)$. □

1.2 Ring of residue classes

Henceforth, assume that $G = \mathbb{Z}/n\mathbb{Z} = \mathbb{Z}_n$, the ring of residue classes modulo n. To distinguish residue classes from integers, specific residue classes will be referred to in boldface, e.g., $\mathbf{0}$, but their variables, e.g., $x \in \mathbb{Z}_p^\times$, are not bolded.

Proposition 1.11 *For SDS in $\mathbb{Z}_n$, there are $\phi(n)$ multiplicative equivalences where ϕ is Euler's totient function.*

Proof By definition there are $\phi(n)$ invertible elements (residue classes d with $gcd(n, d) = 1$) so by Theorem 1.8 there are $\phi(n)$ multiplicative equivalences. □

For p a prime, $\mathbb{Z}_p$ is a field. $(\mathbb{Z}_p)^\times$ includes all $p-1$ non-zero elements and is a cyclic group generated by a **primitive root**, i.e., $(\mathbb{Z}_p)^\times = \{\rho^i : 0 \le x \le p-2\}$ for some ρ whose order is $p-1$. There are $\phi(p-1)$ primitive roots in $\mathbb{Z}_p$.

The following is a well-known fact about finite fields $\mathbb{Z}_p$.

Theorem 1.12 *[Pa1] For $p \equiv 3 \pmod 4$, the set of $\frac{p-1}{2}$ quadratic residues that are perfect squares is an ordinary difference set in $\mathbb{Z}_p$.*

1.3 Cyclotomy

Difference sets exist in many forms, but one way of finding difference sets is through the use of cyclotomial cosets.

Definition 1.13 Let ρ be a primitive root of $(\mathbb{Z}_p)^\times$ where $p = ef + 1$ is prime. Choosing particular e and f, the e-th power **cyclotomial cosets** of $\mathbb{Z}_p$ are:

$$C_i = \{\rho^{es+i} : 0 \le s \le f-1\},\ 0 \le i \le e-1$$

Since $(\mathbb{Z}_p)^\times$ is cyclic, C_0 is the unique subgroup of order f. Hence, the cyclotomial cosets of $\mathbb{Z}_p$ are regular cosets of its multiplicative group, $(\mathbb{Z}_p)^\times$, and as such they partition the non-zero elements of $\mathbb{Z}_p$ into e subsets of f elements.

Cyclotomy has been used to construct difference sets for certain families of prime numbers [Le1]. Two examples are given below of explicit constructions.

Theorem 1.14 *[St1] If $p = 2f + 1 \equiv 3 \pmod 4$, then $S_1 = C_0$ is an ordinary difference set in $\mathbb{Z}_p$, in which case the parameters are $1 - \{p\,;\, \frac{p-1}{2}\,;\, \frac{p-3}{4}\}$.*

Theorem 1.14 is a reformulation of Theorem 1.12 using the language of cyclotomy.

Theorem 1.15 *[Sz2] If $p = 2m + 1$ where $m \equiv 2 \pmod 4$, then $S_1 = C_0 \cup C_1$ and $S_2 = C_0 \cup C_3$ are $2 - \{p\,;\, \frac{p-1}{2}\,;\, \frac{2p-6}{4}\}$ supplementary difference sets in $\mathbb{Z}_p$.*

While $\mathbb{Z}_p$ is additively cyclic, Theorem 1.13 by G. Szekeres also generalizes to additively non-cyclic finite fields, i.e., finite fields of order p^α where $\alpha > 1$.

The problem with these explicit constructions is that only a small number have been found to date, and then they force the parameters to be a certain way. Supplementary difference sets are useful in a number of areas of mathematics, but only with certain restrictions on the parameters that are not met by many known constructions. (These restrictions will be explained in the next section of this paper.) A more ad hoc approach to finding difference sets is via computer search: take a prime and generate the cyclotomial cosets for possible e, then try all combinations of cosets that satisfy given parameter restrictions to see which ones yield supplementary difference sets. This approach is widely employed and various mathematical tools have been developed to facilitate the process.

1.4 Generalized cyclotomy

If n is composite, $\mathbb{Z}_n$ is not a field, and its multiplicative group is not necessarily cyclic. The units of $\mathbb{Z}_n$ are those elements relatively prime to n.

Note that this definition begins with a cyclic subgroup, but there are even more general definitions that allow one to proceed from a non-cyclic subgroup:

Definition 1.16 Take a unit $x \in \mathbb{Z}_n^\times$. Let $\langle x^i \rangle$ be the cyclic group generated by a power of x. Then $y\langle x^i \rangle$ is a **generalized coset** for $y \in \mathbb{Z}_n$.

$$y\langle x^i \rangle = \{yx^{ij} : j = 0, 1, \ldots\}$$

When y is a unit, $y\langle x^i \rangle$ is also a regular coset, in which case it has the same size as $y\langle x^i \rangle$. However, when y is not a unit, then $y\langle x^i \rangle$ can "collapse" into fewer distinct elements. This is because when y is not invertible, it is possible for $ya = yb$ even when $a \neq b$. Ultimately the generalized cosets will partition the elements of $\mathbb{Z}_n$, but into subsets that are not necessarily of the same size.

This definition captures the idea of regular cyclotomy, in which case y is always a unit and all generalized cosets are regular cosets. One minor discrepancy is that in regular cyclotomy the $\{\mathbf{0}\}$ singleton is not considered a coset in the way regular cosets are defined, but here it is considered a generalized coset.

Theorem 1.17 *Multiplication by an invertible element permutes the cosets. Each invertible element can be identified with a bijection of cosets.*

Proof Take invertible d. It is clear that $dy\langle x^i\rangle$ is a coset. Suppose that $dy\langle x^i\rangle = dz\langle x^i\rangle$ for $y\langle x^i\rangle \neq z\langle x^i\rangle$. Since $\exists d^{-1}$, multiplying every element by d^{-1} yields $y\langle x^i\rangle = z\langle x^i\rangle$, a contradiction. So distinct cosets get permuted. □

Theorem 1.18 *In regular cyclotomy, if there are e cosets generated from a primitive root x, multiplication induces e cyclic permutations.*

Proof Take $x^{es+i} \in C_i$, then $x^{es+i}C_j = C_{i+j}$ for all j, so once we know i then we know where all the cosets go, and of course there are e choices of i. □

2 Binary sequences

Definition 2.1 Suppose $a = (a_0, a_1, \ldots, a_{n-1})$ is a complex-valued sequence of length n. Then we define the **periodic autocorrelation function** as:

$$P_a(k) = \sum_{i=0}^{n-1} a_i a_{i+k}$$

where $i + k$ is taken modulo n for every $k = 0, 1, \ldots, n-1$.

Autocorrelation generalizes to a collection of sequences of the same length. If $X = A_1, A_2, \ldots, A_m$ is a collection of sequences of length n, then the periodic autocorrelation function of X is $P_X(k) = \sum_{i=1}^{m} P_{A_i}(k)$. We will be dealing exclusively with pairs of binary-valued sequences, a and b, where $a_i, b_i \in \{\pm 1\}$. For a pair of sequences a and b of length n, $P_{a,b}(k) = \sum_{i=0}^{n-1} a_i a_{i+k} + b_i b_{i+k}$.

Definition 2.2 For a collection of binary sequences X, if there exists a $c \in \mathbb{Z}$ such that $P_X(k) = c$ for every $k \neq 0$, then X is said to be **compatible**.

We are interested in pairs a and b that are compatible with certain values for c. When $c = 0$ and n is even, a and b are called a **periodic Golay pair**. When $c = 2$ and n is odd, a and b are called an **Ehlich pair**. Both are important in the construction of binary matrices of maximal determinant [Dj1] [Gy2].

2.1 Equivalences

There are some operations that preserve the constant autocorrelation function of compatible binary sequences. If pairs of binary sequences are related by shifts or decimations or half-reversal or negation, then they are said to be equivalent.

Definition 2.3 Let $a = (a_i)_{i=0}^{n-1}$ be a sequence. Then $a^l = (a_l, a_{l+1}, \ldots, a_{l+n-1})$ is the **l-th shift** of the sequence a, where the indices are taken modulo n.

Theorem 2.4 *A shift of sequence a preserves $P_a(k)$, so shifting any sequence in a compatible collection preserves the compatibility of the whole collection.*

Proof $\forall l$, $P_{a^l}(k) = \sum_{i=l}^{l+n-1} a_i a_{i+k} = \sum_{i=l}^{n-1} a_i a_{i+k} + \sum_{i=0}^{l-1} a_i a_{i+k} = P_a(k)$. □

Definition 2.5 Let $a = (a_i)_{i=0}^{n-1}$ be a sequence. Taking a_0 and then every d-th entry for some d relatively prime to n is called **decimating** a by d. That is,

$$a_{(d)} = (a_0, a_d, \ldots, a_{(d*(n-1))}) \tag{2}$$

is the d-th decimation of a if $gcd(n, d) = 1$.

It is clear that $gcd(n, d) = 1$ iff $id \not\equiv jd \pmod{n}$ for all $i \not\equiv j \pmod{n}$, so a legitimate decimation rearranges the entries but does not exclude any a_i.

There are $\phi(n)$ decimations of a. Decimation by $-1 \pmod{n}$ is called **reversal**. If $gcd(n, d) = 1$, then $gcd(n, -d \pmod{n}) = 1$. Decimating by d and then performing a reversal is the same as decimating by $-d \pmod{n}$. If we identify each decimation with its reversal, there are $\frac{\phi(n)}{2}$ decimation classes [Fl1].

Definition 2.6 For a sequence a, if we let $p_a = (p_{a_i})_{i=0}^{n-1}$ where $p_{a_i} = P_a(i)$, then p_a is the **autocorrelation sequence** of a.

Lemma 2.7 *If a sequence is decimated by d, then its autocorrelation sequence is also decimated by d.*

Proof $P_{a_{(d)}}(i) = P_a(i * d)$ from (2), and the result follows. □

Theorem 2.8 *If a and b are compatible sequences (they have constant autocorrelation values for $k \neq 0$), then they remain compatible if decimated by the same amount. That is, $a_{(d)}$ and $b_{(d)}$ are compatible for d relatively prime to n.*

The following is a symmetry that exists in the autocorrelation function.

Lemma 2.9 *For any sequence, $P_a(k) = P_a(-k)$.*

Proof $P_a(-k) = \sum_{i=0}^{n-1} a_i a_{i-k} = \sum_{i=-k}^{n-k-1} a_i a_{i+k} = \sum_{i=0}^{n-1} a_i a_{i+k}$ □

Theorem 2.10 *For any sequence, $p_{a_{(d)}} = p_{a_{(-d)}}$. In particular, $p_a = p_{a_{(-1)}}$.*

Proof A reversal exchanges $P_a(k)$ and $P_a(-k)$, $\forall k$. Then apply Lemma 2.9. □

If a and b are compatible then $a_{(-1)}$ and $b_{(-1)}$ are compatible by Theorem 2.8, but also a and $b_{(-1)}$ are compatible, as well as $a_{(-1)}$ and b, by Theorem 2.10. This leads us to the definition of a separate equivalence based on reversal.

Definition 2.11 For a collection of sequences, taking the reversal of a subcollection is called **half-reversal**.

Now the following is an equivalence for binary sequences that, as we shall later see, is special for not having a corresponding equivalence for SDS.

Definition 2.12 Given a sequence, **negation** takes the negative of each entry.

Theorem 2.13 *Negation preserves $P_a(k)$ for any a that is negated, so negating any sequence in a compatible collection preserves compatibility.*

Proof Pairs of 1s that occur k entries apart are swapped for -1s, and vice versa. Similarly, $(1, -1)$ pairs that occur k entries apart are swapped for $(-1, 1)$ pairs. The result is no change to the autocorrelation function for all k. □

2.2 Sequences and difference sets

Binary sequences of length n and subsets of $\mathbb{Z}_n$ are in bijection.

Definition 2.14 Given S a subset of $\mathbb{Z}_n$. If we let $a = (a_i)_{i=0}^{n-1}$ where

$$a_i = \begin{cases} 1 \text{ if } i \in S \\ -1 \text{ if } i \notin S \end{cases}$$

then a is the **incidence sequence** of S. Conversely, we can take the **incidence set** of any binary sequence of length n.

Furthermore, supplementary difference sets and compatible binary sequences are in bijection. First consider the simplest case, that of ordinary difference sets.

Lemma 2.15 *If S is a $1-\{n\,;\,k\,;\,\lambda\}$ ordinary difference set, then its incidence sequence is compatible and*

$$c = n + 4(\lambda - k)$$

Proof Fix $0 < i < n$. Suppose we have a sequence of all -1s, then $P_a(i) = n$. Insert k many 1s into the sequence. Supposing that none of the 1s are i entries apart, each 1 contributes -2 to $P_a(i)$ rather than contributing $+2$ had it remained a -1, so the net change to $P_a(i)$ for each 1 is -4. For each time that a 1 is i entries apart from another 1, this contributes $+2$ to $P_a(i)$ rather than contributing -2 had they not been i entries apart. The incidence sequence of S will have k many 1s and there will be λ many pairs of 1s that are i entries apart, so $P_a(i) = n - 4k + 4\lambda$. Since this holds for all i, the result follows. □

Lemma 2.16 *If a is a compatible sequence, then its incidence set is an ordinary difference set.*

Proof Letting $\alpha = \sum_{i=0}^{n-1} a_i$, there are $\frac{n+\alpha}{2}$ 1s in this sequence of length n. With reference to the proof of Lemma 2.15, the only way that $P_a(i)$ is the same for all i, is that the 1s occur i entries apart the same number of times for each i. But then this means that the incidence set for a is an ordinary difference set, with $k = \frac{n+\alpha}{2}$. Now we can apply 2.15 and solve for λ to get $\lambda = k + \frac{c-n}{4}$. □

Lemmas 2.15 and 2.16 can be generalized from ordinary difference sets to all supplementary difference sets with essentially the same proofs, adapted for multiple sequences and sets instead of just one. There is a bijection between SDS and compatible sequences and they are related by the following theorems.

Theorem 2.17 *Let $S_1, S_2, \ldots, S_e$ be $e - \{n\,;\,k_1, k_2, \ldots, k_e\,;\,\lambda\}$ be SDS in $\mathbb{Z}_n$. Then their incidence sequences are compatible with*

$$c = en + 4\left(\lambda - \sum_{i=1}^{e} k_i\right)$$

Theorem 2.18 *Let $A_1, A_2, \ldots, A_e$ be compatible sequences of length n. Let $\alpha_1, \alpha_2, \ldots, \alpha_e$ be the entry sums. Then their incidence sets, $S_1, S_2, \ldots, S_e$, are $e - \{n\,;\, k_1, k_2, \ldots, k_e\,;\, \lambda\}$ SDS with*

$$k_i = \frac{n + \alpha_i}{2}, \qquad \lambda = \sum_{i=1}^{e} k_i + \frac{c - en}{4}$$

In the case we are interested in, SDS pairs, we have the following statements.

Proposition 2.19 *[Gy2] Let a and b be compatible sequences. Let α and β be the entry sums. Their incidence sets, S_a and S_b, are $2 - \{n\,;\, k_a, k_b\,;\, \lambda\}$ SDS with*

$$k_a = \frac{n + \alpha}{2}, \qquad k_b = \frac{n + \beta}{2}, \qquad \lambda = k_a + k_b + \frac{c - 2n}{4}$$

Proposition 2.20 *Let S_a and S_b be SDS with parameters $2 - \{n\,;\, k_a, k_b\,;\, \lambda\}$. Then their incidence sequences, a and b, are compatible and*

$$\alpha = 2k_a - n, \qquad \beta = 2k_b - n, \qquad c = 2n + 4(\lambda - k_a - k_b)$$

The upshot is that finding compatible sequence pairs with specific c is equivalent to finding supplementary difference sets with the requisite parameters.

Now we can relate the equivalences of sequences (shifts, decimations, half-reversal) with those for difference sets (addition, multplication, half-multiplication).

Theorem 2.21 *Taking the l-th shift of a sequence corresponds to subtracting l from its incidence set.*

Proof Take $a^l = (a_l, a_{l+1}, \ldots, a_{l+n-1})$. Since $i \in S_a \Rightarrow a_i = 1$, and this a_i is shifted to the $i - l$-th position of a^l, this means that $i \in S_a \Rightarrow i - l \in S_{a^l}$. □

Theorem 2.22 *Taking the d-th decimation of a sequence (which implies that d is invertible) corresponds to multiplying its incidence set by d^{-1}.*

Proof Take $a_{(d)} = (a_0, a_d, \ldots, a_{(d*(n-1))})$. Decimation by d moves a_i to the position x where $i = dx$, and so a_i is moved to the $d^{-1}i$-th position of $a_{(d)}$. This means that $i \in S_a \Rightarrow a_i = 1 \Rightarrow d^{-1}i$-th position of $a_{(d)}$ is $1 \Rightarrow d^{-1}i \in S_{a_{(d)}}$. □

By our discussion on decimations, half-reversal of a collection of sequences corresponds to multiplying incidence sets by $\mathbf{-1}^{-1} = \mathbf{-1}$, or half-multiplication. This is an equivalence of supplementary difference sets by Theorem 1.10.

We see that some operations on sequences that preserve compatibility have analogous operations that preserve both SDS and parameters of their incidence sets. And those operations on sets that preserve SDS and parameters have analogous operations that preserve the compatibility of their incidence sequences.

Negation of sequences preserves compatibility by Theorem 2.13, and so by the discussion above it must also preserve the SDS of the incidence sets. But taking the negation of some sequences corresponds to taking the complement of their incidence sets, which although preserving SDS, does not preserve the parameters k_i and λ. So even though negation qualifies as an equivalence for sequences, its corresponding action is not an equivalence for difference sets.

2.3 Sequences and cyclotomy

The following is a sum-of-squares condition for compatible binary sequence pairs. It is one way to narrow down the possibilities of coset combinations when searching for difference sets using cyclotomy.

Theorem 2.23 *[Gy2] Suppose a and b are compatible binary sequences of length n so that $P_a(k) + P_b(k) = c$ for all $k \neq 0$. Let α be the entry sum of a and β the entry sum of b. That is, $\alpha = \sum_{i=0}^{n-1} a_i$ and $\beta = \sum_{i=0}^{n-1} b_i$. Then*

$$\alpha^2 + \beta^2 = 2n + cn - c \tag{3}$$

Proof

$$\alpha^2 + \beta^2 = \left(\sum_{i=0}^{n-1} a_i\right)^2 + \left(\sum_{i=0}^{n-1} b_i\right)^2 = \sum_{k=0}^{n-1} \left(P_a(k) + P_b(k)\right),$$

$$\sum_{k=0}^{n-1} \left(P_a(k) + P_b(k)\right) = 2n + \sum_{k=1}^{n-1} \left(P_a(k) + P_b(k)\right) = 2n + (n-1)c$$

□

Finding an Ehlich pair ($c = 2$) of length n requires that $4n - 2$ be a sum of two squares. Let $n = 19$. Then $4n-2 = 7^2+5^2$ In some cases there are multiple representations as a sum of two squares. Here there is only one, so $\alpha = \pm 7$ and $\beta = \pm 5$. Suppose $\alpha = 7$ and $\beta = 5$; then there are 13 1s in sequence a and 12 1s in sequence b. Looking at $\mathbb{Z}_{19}$, if we let $e = 3$ and $f = 6$, then we need two cosets and $\{0\}$ for S_a, and two cosets for S_b. This does not guarantee that an SDS exists, but a computer search would try coset combinations of this form.

In the periodic Golay case ($c = 0$), $\alpha^2 + \beta^2 = 2n$, which implies that n is also a sum of two squares. Already this narrows down the possibilities for n. From number theory, we know that $n \in \mathbb{Z}_+$ is a sum of two squares if and only if all of its prime divisors that are congruent to 3 (mod 4) have an even power.

In 2008, Adam Vollrath proved the following about periodic Golay pairs.

Theorem 2.24 *[Vo1] Assume a and b form a periodic Golay pair of length $n = 2m$ for some odd m, and assume that there is a unique representation of n*

as a sum of two squares. Let E_a be the sum of the even entries of a and D_a be the sum of the odd entries of a, and similarly for E_b and D_b. Then

$$n = E_a^2 + D_a^2 = E_b^2 + D_b^2$$

Regular cyclotomy cannot be used to find periodic Golay pairs because of the condition that n is even. The next two sections of this paper will be focused on finding periodic Golay pairs for $n = 2p$ using generalized cyclotomy.

3 Generalized cyclotomy for $2p$

First we make some remarks on notation. Brackets around an integer ξ signifies taking its residue class, so $[\xi]_{2p}$ is the residue class of ξ modulo $2p$. In this section, *order* always means multiplicative order.

Definition 3.1 Take $x \in \mathbb{Z}_n$. If, for every integer $\xi \in x$, $\xi \equiv k \pmod{n}$ for $k < n$ implies that k is an even integer, then x is called an **even residue class**.

Definition 3.2 Take $x \in \mathbb{Z}_n$. If, for every integer $\xi \in x$, $\xi \equiv k \pmod{n}$ for $k < n$ implies that k is an odd integer, then x is called an **odd residue class**.

Note that an integer ξ can belong to an even residue class of $\mathbb{Z}_n$ without being an even integer, or to an odd residue class without being an odd integer.

3.1 Coset structure

In this section we consider $n = 2p$ for p a prime. $\mathbb{Z}_{2p}^{\times}$ is cyclic by the Chinese Remainder Theorem, and consists of the $p-1$ odd residue classes besides $\mathbf{p}$.

If $p = ef + 1$, we let C_0 be the unique cyclic subgroup of order f. That is, $C_0 = \{\rho^{es} : 0 \leq s \leq f-1\}$ for primitive $\rho \in \mathbb{Z}_{2p}^{\times}$. If we let $C_i = \rho^i C_0$, $i \leq e-1$, we have e regular cosets of size f which partition the non-$\mathbf{p}$ odd residue classes.

Now we examine the generalized cosets, which consist of the even residues.

Lemma 3.3 *If $\mathbf{2}[\xi]_{2p} = \mathbf{0} \in \mathbb{Z}_{2p}$, then $[\xi]_{2p} = \mathbf{0}$ or $\mathbf{p}$.*

Proof By assumption, $(2pk + 2)(\xi) = 2pl$ for some k, l, which requires that $2\xi = 2pm$ for some m, and this implies that $\xi \equiv p \pmod{2p}$ or $0 \pmod{2p}$. □

Lemma 3.4 *If $x, y \in \mathbb{Z}_{2p}^{\times}$ and $\mathbf{2}x = \mathbf{2}y$, then $x = y$.*

Proof Since $\mathbf{2}(x-y) = \mathbf{0}$, by Lemma 3.3 $(x-y) = \mathbf{0}$ or $\mathbf{p}$. If $(x-y) = \mathbf{p}$, then either x or y is even, which contradicts that both elements belong in $\mathbb{Z}_{2p}^{\times}$. □

Residues modulo $2p$ follow the usual even/odd multiplication rules. If x is an even and y an odd residue, then xy is an even residue. As a result, if $x \in \mathbb{Z}_{2p}$ is even while $y \in \mathbb{Z}_{2p}^{\times}$ is odd, then $xy \notin \mathbb{Z}_{2p}^{\times}$ while $xy + z \in \mathbb{Z}_{2p}^{\times}$ for odd z.

Theorem 3.5 *Assume $p = ef + 1$. If the regular cosets partition the non-$\mathbf{p}$ odd residues into e sets of size f, then some generalized cosets partition the non-$\mathbf{0}$ even residues into e sets of size f. They are of the form $\mathbf{2}C_0, \mathbf{2}C_1, \ldots, \mathbf{2}C_{e-1}$.*

Proof Multiply each C_i by $\mathbf{2}$ and apply Lemma 3.4 to get $p-1$ distinct even residues. They are non-zero since elements of $\mathbb{Z}_{2p}^{\times}$ cannot be zero divisors. □

Finally, it is easy to see see that $\mathbf{p}C_0 = \{\mathbf{p}\}$ and $\mathbf{0}C_0 = \{\mathbf{0}\}$, so in total we have $2e$ (regular or generalized) cosets of size f plus these two singletons.

From here on we will refer to $\mathbf{2}C_i$ as $2C_i$.

3.2 Relation to $\mathbb{Z}_p^{\times}$

Definition 3.6 Let f be the **match** function $f : \mathbb{Z}_{2p}^{\times} \to \mathbb{Z}_p^{\times}$

$$f([\xi]_{2p}) = [\xi]_p$$

Proposition 3.7 *f is bijective.*

Proof Each $[\xi]_{2p} < [p]_{2p}$ goes to a distinct odd residue of $\mathbb{Z}_p^{\times}$ and each $[\xi]_{2p} > [p]_{2p}$ goes to a distinct even residue, accounting for all $p-1$ residues in $\mathbb{Z}_p^{\times}$. □

The inverse is

$$f^{-1}([\xi]_p) = \begin{cases} [\xi]_{2p} & \text{if } \xi \pmod p \text{ is an odd residue} \\ [\xi + p]_{2p} & \text{if } \xi \pmod p \text{ is an even residue} \end{cases}$$

Proposition 3.8 *If $\xi \pmod{2p}$ is an odd residue, then ξ is an odd integer.*

Proof If $\xi = 2pk + \eta$ for some k and some odd η, then clearly ξ is odd. □

Theorem 3.9 *If $[\xi]_{2p} \in \mathbb{Z}_{2p}^{\times}$ has order d, then $[\xi]_p$ has order d.*

Proof By assumption, $\xi^d = 2pk+1$ for some k. Then certainly $\xi^d \equiv 1 \pmod p$. Suppose, for some $e < d$, $\xi^e = pf+1$ for some f. Because ξ is odd by Prop. 3.8, ξ^e is odd, so f must be even, so $\xi^e = 2p(f/2)+1$ and $\xi^e \equiv 1 \pmod{2p}$, which contradicts d being the order of $\xi \pmod{2p}$. Thus d is the order of $\xi \pmod p$. □

In particular, $\rho \in \mathbb{Z}_{2p}^{\times}$ is a primitive root iff $f(\rho)$ is primitive in $\mathbb{Z}_p^{\times}$.

Proposition 3.10 *For $x, y \in \mathbb{Z}_{2p}^{\times}$, $f(xy) = f(x)f(y)$.*

Proof $f([\xi]_{2p}[\gamma]_{2p}) = f([\xi\gamma]_{2p}) = [\xi\gamma]_p = [\xi]_p[\gamma]_p = f(\xi)f(\gamma)$ □

If $f(\rho^i) = f(\rho)^i$ for all i, then it follows that the cyclotomial cosets are related in the same manner. Denote $\{f(x) : x \in C_i\}$ by $f(C_i)$. If $(C_i)_{i=0}^{e-1}$ are the regular cosets in $\mathbb{Z}_{2p}^{\times}$ generated by ρ, then $(f(C_i))_{i=0}^{e-1}$ are the cosets in $\mathbb{Z}_p^{\times}$ generated by $f(\rho)$. Conversely, if $(C_i)_{i=0}^{e-1}$ are the cosets generated by primitive $\rho \in \mathbb{Z}_p^{\times}$, then $(f^{-1}(C_i))_{i=0}^{e-1}$ are the regular cosets in $\mathbb{Z}_{2p}^{\times}$ generated by $f^{-1}(\rho)$.

3.3 Equivalences and coset structure

In order to develop a classification of SDS constructed from cyclotomial cosets, we need to understand the equivalences that preserve coset structure.

Theorem 3.11 *Multiplication by invertible x induces a cyclic permutation of the regular cosets, which permutes the generalized cosets in the same way.*

Proof If $x \in C_i$, then $x = \rho^{es+i}$ for some s. For any $y \in C_j$, $y = \rho^{et+j}$ for some t, so $xy = \rho^{e(s+t)+(j+i)}$, so $xC_i = C_{i+j}$ where the index $i+j$ is taken modulo e. If $xC_i = C_{i+j}$, then $2xC_i = 2C_{i+j}$. Finally, $x\{p\} = \{p\}$ and $x\{0\} = \{0\}$. □

Theorem 3.12 *If $xC_i = C_{i+j}$ in $\mathbb{Z}_{2p}^{\times}$, then $f(x)f(C_i) = f(C_{i+j})$ in $\mathbb{Z}_p^{\times}$.*

Proof Assuming that the cosets in $\mathbb{Z}_{2p}^{\times}$ are generated by ρ and the cosets in $\mathbb{Z}_p^{\times}$ are generated by $f(\rho)$, recall from the discussion above that $f(C_i)$ are the cosets of $\mathbb{Z}_p^{\times}$. If we have $xC_i = C_{i+j}$, then $f(C_{i+j}) = f(xC_i) = f(x)f(C_i)$. □

If we know what happens to the cosets under multiplication in $\mathbb{Z}_p^{\times}$, then we can formulate explicitly what happens to the cosets of $\mathbb{Z}_{2p}^{\times}$, and vice versa.

Theorem 3.13 *Addition by $\mathbf{p}$ preserves coset structure by swapping regular and generalized cosets.*

Proof Let $[2\xi]_{2p} \in 2C_i$, so $[2\xi]_{2p}$ is even. We have $[2\xi + p]_{2p} \in 2C_i + \mathbf{p}$ and then $[2\xi + p]_{2p} \in \mathbb{Z}_{2p}^{\times}$ so we can take $f([2\xi + p]_{2p}) = [2\xi + p]_p = [2\xi]_p = [2]_p f([\xi]_{2p})$. Since $[\xi]_{2p} \in C_i$, what this means is that $2C_i + \mathbf{p} = f^{-1}([2]_p)C_i$. Multiplying both sides, we have $(f^{-1}([2]_p))^{-1}(2C_i + \mathbf{p}) = C_i$, and $(f^{-1}([2]_p))^{-1}2C_i + \mathbf{p} = C_i$, so $C_i + \mathbf{p} = (f^{-1}([2]_p))^{-1}2C_i$. Finally, $\{p\} + \mathbf{p} = \{0\}$ and $\{0\} + \mathbf{p} = \{p\}$. □

Corollary 3.14 *If $[2]_p \in f(C_j)$, then $C_i + \mathbf{p} = 2C_{i-j}$ and $2C_i + \mathbf{p} = C_{i+j}$.*

Proof If $[2]_p \in f(C_j)$, then $f^{-1}([2]_p) \in C_j$ and $(f^{-1}([2]_p))^{-1} \in C_{-j}$. Referring to the proof of Theorem 3.13, $C_i + \mathbf{p} = (f^{-1}([2]_p))^{-1}2C_i$, which implies that $C_i + \mathbf{p} = 2C_{i-j}$, and $2C_i + \mathbf{p} = f^{-1}([2]_p)C_i$, which implies that $2C_i + \mathbf{p} = C_{i+j}$. □

If we know to which coset $\mathbf{2}$ belongs in $\mathbb{Z}_p$, then we can write out the coset permutation when adding by $\mathbf{p}$ in $\mathbb{Z}_{2p}$. We do so in the next sub-section.

Suppose we added a non-zero element other than $\mathbf{p}$, say $x = [\gamma]_{2p}$. If x is even, then $C_i + x$ remains in $\mathbb{Z}_{2p}^{\times}$, so we can take $f(C_i + x)$. Given $[\xi]_{2p} + x \in C_i + x$, we have $f([\xi]_{2p} + x) = [\xi + \gamma]_p$, but there is no reason to believe that the union of all such $[\xi + \gamma]_p$ equals a coset in $\mathbb{Z}_p$. The same argument applies to $f(2C_i + x)$ if x were odd. In fact, there are no more elements that preserve coset structure additively because the $\mathbf{p}$-addition is the only one that swaps the singletons, $\{\mathbf{p}\}$ and $\{\mathbf{0}\}$. So while all the invertible elements respect cosets by multiplication, only the element $\mathbf{p}$ does for addition. From the perspective of sequences, all decimations respect cosets but only the p-th shift does.

3.4 Cyclotomy formulas for equivalent SDS

Fix e and f where $p = ef + 1$. If we have a pair of supplementary difference sets in $\mathbb{Z}_{2p}$ that are constructed from cyclotomial cosets, they will be of the form:

$$S_a = \bigcup_{i \in I} C_i \bigcup_{j \in J} 2C_j \qquad S_b = \bigcup_{m \in M} C_m \bigcup_{n \in N} 2C_n \tag{4}$$

without taking into account the singletons, $\{\mathbf{0}\}$ and $\{\mathbf{p}\}$.

We are interested in finding formulas for equivalent SDS that also respect coset structure. By Theorem 3.11, multiplication by $x \in C_k$ is an equivalence that cyclically permutes the cosets by k. There are e choices for k, including $k = 0$. So multiplication gives us e equivalent SDS pairs that respect coset structure. From the perspective of cyclotomy, it does not matter which $x_k \in C_k$ we choose since each representative will permute the cosets in exactly the same way.

Recall that multiplication is an equivalence if we multiply both sets by the same element. Suppose we are multiplying by $x_k \in C_k$. Recall that multiplying by x_i corresponds to decimating the incidence sequences by x_k^{-1}. Thus we have:

$$S_{a_{\left(x_k^{-1}\right)}} = \bigcup_{i \in I} C_{i+k} \bigcup_{j \in J} 2C_{j+k} \quad S_{b_{\left(x_k^{-1}\right)}} = \bigcup_{m \in M} C_{m+k} \bigcup_{n \in N} 2C_{n+k}, \quad 0 \le k < e \tag{5}$$

with no change to the singletons.

Now suppose we perform half-multiplication. We know from Storer (Lemma 2, pg. 24) that if f is even then $-\mathbf{1} \in C_0$ and if f is odd then e is even and $-\mathbf{1} \in C_{\frac{e}{2}}$ [St1]. In the former case, the cyclic permutation induced on the cosets is identity, so we do not get new SDS. In the latter case, however, we get new coset combinations by a cyclic permutation of $e/2$, and we show this as:

$$S_a = \bigcup_{i \in I} C_i \bigcup_{j \in J} 2C_j \qquad S_{b_{(-1)}} = \bigcup_{m \in M} C_{m+\frac{e}{2}} \bigcup_{n \in N} 2C_{n+\frac{e}{2}} \tag{6}$$

with no change to the singletons. Note that we arbitrarily took a and $b_{(-1)}$ when we could have taken $a_{(-1)}$ and b.

Recall that when adding $\mathbf{p}$, the corresponding action on the sequence is taking the $(-p)$-th shift. For adding $\mathbf{p}$, an application of Corollary 3.14 gives:

$$S_{a^{-p}} = \bigcup_{j \in J} C_{j+s} \bigcup_{i \in I} 2C_{i-s} \qquad S_b = \bigcup_{m \in M} C_m \bigcup_{n \in N} 2C_n \tag{7}$$

where $f(\rho)^{et+s} = \mathbf{2} \in \mathbb{Z}_p$ for some t, or in other words, $[2]_p \in f(C_s)$. Note that we arbitrarily took a^p and b when we could have taken a and b^p or a^p and b^p.

The singletons (if any were included in S_a, S_b) are swapped with each other.

So far we have taken individual operations on the original SDS, but in order to map out an equivalence class we have to consider multiple operations in sequence. In the $2p = 26$ case, a computer program written by Adam Vollrath and Dr. William Orrick (2008) found 32 SDS using generalized cyclotomy (with $e = 4$ and $f = 3$ and $\rho = \mathbf{7}$) that give periodic Golay pairs. It turns out that all 32 are equivalent pairs according to our classification scheme. Take one pair and there are 4 distinct decimations. Then perform half-reversals on each of the 4 and we have 8 distinct pairs. Then there are four possibilities for **p**-shifts, and so we have 32 in total. (Neither the computer program nor our classification scheme has accounted for complementation, which is easy enough to figure out and work with by hand.) The $2p = 82$ case is similar except instead of one equivalence class, we have 9 equivalence classes. In the 82 case, $e = 8$ and $f = 5$, so there are 8 distinct decimation classes of an initial pair within a given sequence, and then half-reversals give us 16, from which we get 64 pairs by shifts. Then multiply by 9 equivalence classes to get 576 SDS pairs, which is the number the computer program found. Since the number of solutions was so large, Dr. Orrick used Mathematica to check that these equivalence classes do exist among the solutions and that they account for all the solutions.

3.5 Cosets and repetition numbers

If we let $S = \bigcup_{i\in I} C_i \bigcup_{j\in J} 2C_j$ and $x \in \mathbb{Z}_{2p}$, then we call λ_x the repetition number of x if x occurs λ_x times in ΔS. First we show that λ_x respects cosets.

Theorem 3.15 *If $x, y \in C_i$, then $\lambda_x = \lambda_y$.*

Proof If $\rho^{es+n} \in C_n \subset S$ and $\rho^{et+m} \in C_m \subset S$ such that $\rho^{es+n} - \rho^{et+m} = \rho^{er+i} \in C_i$, we have $\rho^{e(s+k)+n} - \rho^{e(t+k)+m} = \rho^{e(r+k)+i} \in C_i$ for $0 \le k < f$. Each element of C_i occurs once on the right-hand side. Since $\rho^{e(s+k)+n} \in C_n$ and $\rho^{e(t+k)+m} \in C_m$ for all k, each of the elements of C_i occurs in ΔS. So whenever an element of C_i occurs in ΔS, the other elements of C_i occur in ΔS as well. □

Now that we have established that repetition numbers are the same within each coset, let $\lambda(C_i)$ denote the repetition number for all elements in C_i.

Proposition 3.16 *If f is even, then $2 \mid \lambda(C_i)$ for all i.*

Proof If f is even then $-\mathbf{1} \in C_0$. If $x - y \in \Delta S$ then $y - x \in \Delta S$. If $x - y \in C_i$ then $y - x \in -C_i = C_i$, so each pair $x - y \in C_i$ contributes 2 to $\lambda(C_i)$. □

Proposition 3.17 *If f is odd, then $\lambda(C_i) = \lambda(C_{i+\frac{e}{2}})$.*

Proof If f is odd then $-\mathbf{1} \in C_{\frac{e}{2}}$. If $x - y \in \Delta S$ then $y - x \in \Delta S$. If $x - y \in C_i$ then $y - x \in -C_i = C_{i+\frac{e}{2}}$, so each pair that adds to $\lambda(C_i)$ also adds to $\lambda(C_{i+\frac{e}{2}})$. □

Since **p** belongs in its own coset, we consider the special case of $\lambda_{\mathbf{p}}$.

Theorem 3.18 *$\lambda_{\mathbf{p}} = 2f\theta$ where θ is the number of coset pairs in S of the form $(C_i, 2C_{i-j})$, where j is such that $[2]_p \in f(C_j)$. As a result, $2f \mid \lambda_{\mathbf{p}}$.*

Proof By Corollary 3.14, $C_i + \mathbf{p} = 2C_{i-j}$, so there are f many $x \in C_i$, $y \in 2C_{i-j}$ such that $x + \mathbf{p} = y$, which means $x - y = -\mathbf{p} = \mathbf{p}$. For each of those choices, $y - x = \mathbf{p}$, so each pair of cosets contributes $2f$ to $\lambda_{\mathbf{p}}$. Conversely, if $x \in C_m$ and $y \in 2C_n$ are such that $x - y = \mathbf{p}$, then $\rho^{ek}x - \rho^{ek}y = \rho^{ek}\mathbf{p} = \mathbf{p}$ for $0 \leq k < f$, so then $C_m + \mathbf{p} = 2C_n$ By Corollary 3.14, n must be of the form $m - j$. □

Now we see what happens with the repetition numbers when we introduce the singletons. Denote $S_0 = S \cup \{\mathbf{0}\}$, $S_p = S \cup \{\mathbf{p}\}$, and $S_{0,p} = S \cup \{\mathbf{0}\} \cup \{\mathbf{p}\}$.

Proposition 3.19 *If $C_i \subset S$, then $\lambda(C_i)$ and $\lambda(-C_i)$ increase by 1 in S_0.*

Proof For each $C_i \in S$, in ΔS_0 we have $C_i - \mathbf{0} = C_i$ and $\mathbf{0} - C_i = -C_i$. □

Proposition 3.20 *If $C_i \subset S$, then $\lambda(2C_{i-j})$ increases by 2 in S_p, and if $2C_i \subset S$, then $\lambda(C_{i+j})$ increases by 2 in S_p, where j is such that $[2]_p \in f(C_j)$.*

Proof This follows from Corollary 3.14 and the fact that $\mathbf{p} = -\mathbf{p}$. □

Proposition 3.21 *In addition to the changes recorded in Prop. 3.19 and Prop. 3.20 from including $\{\mathbf{0}\}$ and $\{\mathbf{p}\}$ separately, taking $S_{0,p}$ will increase $\lambda_{\mathbf{p}}$ by 2.*

Proof The only differences left to be accounted for are those between $\mathbf{p}$ and $\mathbf{0}$, in which case we have $\mathbf{p} - \mathbf{0} = \mathbf{p}$ and $\mathbf{0} - \mathbf{p} = \mathbf{p}$, so $\lambda_{\mathbf{p}}$ increases by 2. □

3.6 Parameter and coset constraints for SDS

An important fact from the preceding discussion is that $\lambda_{\mathbf{p}} \equiv 0 \pmod{2f}$ if a cyclotomy-constructed set does not contain both of the singletons, and $\lambda_{\mathbf{p}} \equiv 2 \pmod{2f}$ if a cyclotomy-constructed set does contain both of the singletons. Thus we have the following theorem which gives us a useful constraint on the parameters of SDS, and which can be used to rule out certain sets of parameters.

Theorem 3.22 *For SDS pairs S_a and S_b in $\mathbb{Z}_{2p}$, as long as $f > 2$:*
$\lambda \equiv 0 \pmod{2f}$ *if both* k_a *and* $k_b \equiv 0$ *or* $1 \pmod{f}$
$\lambda \equiv 2 \pmod{2f}$ *if* $k_a \equiv 0$ *or* $1 \pmod{f}$ *and* $k_b \equiv 2 \pmod{f}$
$\lambda \equiv 4 \pmod{2f}$ *if both* k_a *and* $k_b \equiv 2 \pmod{f}$

Proof The first case corresponds to both sets being constructed from cyclotomial cosets but neither having both of the singletons, so $\lambda_{\mathbf{p}}$ can only be a multiple of $2f$, so the overall λ has to be a multiple of $2f$ as well. The second case corresponds to one set having both singletons and the other does not, so $\lambda_{\mathbf{p}}$ is a multiple of $2f$ plus 2 from the two singletons, and so λ has to be that as well. Finally the third case corresponds to both sets having both singletons. □

These constraints can be used to rule out parameter sets of SDS pairs. This can be done to speed up computer searches when dealing with large values of n,

in which case there are large numbers of parameter sets to try, and trying each parameter set takes a good deal of time to generate all the coset combinations.

The following is a constraint on the coset combinations that an SDS pair must have.

Theorem 3.23 *Let S_a and S_b be SDS pairs in $\mathbb{Z}_{2p}$ with $f > 2$. If $2f$ divides into λ k times (with or without a remainder) then, between the two sets, there are exactly k coset pairs of the form $(C_i, 2C_{i-j})$, where j is such that $[2]_p \in f(C_j)$.*

Proof This is a straightforward consequence of Theorem 3.18. □

So, for given $n = 2p$, Theorem 3.22 narrows down the feasible parameter sets, and once we know a parameter set is viable, Theorem 3.23 can tell us exactly how many coset pairs go into the construction of an SDS pair with those parameters, and just as importantly, that no more than that number of coset pairs of that form can go into the construction. This improves computer searches when the number of cosets is very large.

References

[Bl1] D. Blatt, G. Szekeres: A skew Hadamard matrix of order 52. Canadian J. Math. 21 (1969), 1319–1322.

[Dj1] D. Djoković: On Maximal (1, -1)-Matrices of Order 2n, n Odd. Radovi Matematički Vol. 7 (1991), 371–378.

[Fl1] R.J. Fletcher, M. Gysin, J. Seberry: Application of the Discrete Fourier Transform to the Search for Generalised Legendre Pairs and Hadamard Matrices. Australasian Journal of Combinatorics, 23 (2001), 75–86.

[Gy1] M. Gysin, J. Seberry: An Experimental Search and New Combinatorial Designs via a Generalisation of Cyclotomy. JCMCC 27 (1998), 143–160.

[Gy2] M. Gysin: New D-Optimal Designs via Cyclotomy and Generalised Cyclotomy. Australasian Journal of Combinatorics, 15 (1997), 247–255.

[Ko1] C. Koukouvinos, S. Kounias, J. Seberry: Supplementary difference sets and optimal designs. Discrete Mathematics 88 (1991), 49–58.

[Ko2] S. Kounias, C. Koukouvinos, N. Nikolaou, A. Kakos: The Non-equivalent Circulant D-Optimal Designs for $n \equiv 2 \bmod 4$, $n \leq 54$, $n = 66$. Journal of Combinatorial Theory, Series A 65 (1994), 26–38.

[Le1] E. Lehmer: A family of supplementary difference sets. Bull. Austral. Math. Soc. Vol. II (1974), 1–4.

[Pa1] R. Paley: On orthogonal matrices. J. Math. Phys., 2. (1933), 311-320.

[St1] T. Storer: Cyclotomy and Difference Sets. Lectures in Advanced Mathematics, 2. Markham Publishing Company, Chicago, 1967.

[Sz1] G. Szekeres: Cyclotomy and complementary difference sets. Acta Arithmetica XVIII (1971), 349–353.

[Sz2] G. Szekeres: Tournaments and Hadamard matrices. L'Enseignement Math. 15 (1969), 269–278.

[Vo1] A. Vollrath: A modification of periodic Golay sequence pairs. Mathematics Dept., Indiana University (2008).

Smale's Mean Value Conjecture with Complex Dynamics for Quartic Polynomials

Nicholas Miller and Max Zhou

Abstract

We will discuss a version of the Smale Mean Value Conjecture (SMVC) that includes a condition on the dynamics of the critical points of complex polynomials. We call this variant of the SMVC the Dynamical Smale Mean Value Conjecture (DSMVC). Our main results are proofs of the DSMVC for quartics with repeated critical points and quartics with all critical points real. Finally, we will discuss partial results on when general quartic polynomials satisfy the DSMVC.

1 Introduction

1.1 Smale's Mean Value Conjecture

In 1981, Stephen Smale posed the following problem in [**?**]:

Conjecture 1.1 (Smale's Mean Value Conjecture) *Let P be a non-linear, complex polynomial of degree d, with critical points c_i and $P'(0) \neq 0$. If $z \in \mathbb{C}$ is not a critical point of P, then does*

$$\min_i \left| \frac{P(z) - P(c_i)}{z - c_i} \right| \leq K \left| P'(z) \right|. \tag{1}$$

hold for $K = 1$ or even $K = \frac{d-1}{d}$?

Smale noted that $K = \frac{d-1}{d}$ would be the sharpest possible bound if true, as the bound is attained for $\tilde{P}(z) = z^d - z$ and taking $z = 0$ in (1).

While progress has been made, Conjecture 1.1 is still open in full generality. Specifically, in [**?**], Conjecture 1.1 was proved for $d \leq 10$; however, little is known for $d > 10$, except for special cases. A brief survey of results is presented in [**?**].

1.2 Dynamical Smale Mean Value Conjecture

In [**?**], Pilgrim and Miles-Leighton formulated a stronger conjecture that imposed a condition on the dynamics of critical points:

Conjecture 1.2 (DSMVC) *Let $f(z) = z + a_2z^2 + \cdots + a_dz^d$ be a non-linear complex polynomial. Then, there exists a critical point c of f for which $\left|\frac{f(c)}{c}\right| \leq 1$ and which in addition converges to the origin under iteration of f.*

Pilgrim and Miles-Leighton proved Conjecture 1.2 for $d = 2, 3$. In this report, we discuss partial results on this conjecture for $d = 4$.

1.3 Our Investigations

For the rest of the report, the bound $\left|\frac{f(c)}{c}\right| \leq 1$ will be called 'the SMVC' and Conjecture 1.2 will be called 'the DSMVC.' Additionally, we will assume all polynomials are over $\mathbb{C}$.

Our investigations have dealt mainly with the DSMVC. The SMVC states that after 1 iteration of the polynomial, there exists a critical point that moves closer to the origin. As will be discussed later, Theorem 2.8 states for any complex polynomial of the form as in the DSMVC hypothesis, there also exists a critical point converging to the origin under iteration. For $d \leq 10$, we know there exists a critical point satisfying the SMVC. For a given polynomial, do the same critical points satisfy the SMVC and converge to the origin under iteration? Does one of the conditions imply the other? Will examining the dynamics of critical points give more insight into which critical points satisfy the SMVC?

2 Complex Dynamics

Before we present our results, we will briefly describe the field of complex dynamics, and present some relevant definitions/results.

The field of complex dynamics studies how dynamical systems over $\mathbb{C}$ behave. A dynamical system is a system whose state changes over time deterministically, and can take values from the phase space. In our case, the phase space is $\mathbb{C}$, the system is points in $\mathbb{C}$, the change of state is determined by functional iteration, and time is measured in integer values based on the specific functional iterate.

To make the discussion more precise, we will introduce some basic definitions/notations:

Definition 2.1 We will denote $f^n(p)$ as the n-fold composition of the function f, evaluated at p, where $n \in \mathbb{N}$. Define $f^0(p) = p$.

Definition 2.2 An ***orbit*** of a point p under the function f is the sequence of points $\{f^n(p)\}_{n\in\mathbb{N}}$.

There is a particular type of point defined by its orbit:

Definition 2.3 A point p is a ***periodic point*** of period $k \in \mathbb{Z}^+$ under a function f if $f^k(p) = p$.

A special case of a periodic point is:

Definition 2.4 A point p is a ***fixed point*** under a function f if p is a periodic point of period 1, that is: $f(p) = p$.

An essential equivalence relation of dynamical systems that we will use often is:

Definition 2.5 Functions f and g are ***conjugate*** if there exists an invertible function h such that $f = h \circ g \circ h^{-1}$. We say "$f$ is conjugate to g" or "conjugate g by h to produce f."

Important properties of conjugacy are that it preserves fixed points and critical points (where the derivative vanishes):

Lemma 2.6 *If $f = h \circ g \circ h^{-1}$ and p is a fixed point of g, then $h(p)$ is a fixed point of f.*

Proof It is easy to see that $f(h(p)) = h \circ g \circ h^{-1}(h(p)) = h \circ g(p) = h(p)$. □

Lemma 2.7 *If $f = h \circ g \circ h^{-1}$ and c is a critical point of g, then $h(c)$ is a critical point of f.*

Proof From the chain rule, $f'(h(c)) = h'(g \circ h^{-1}(h(c))) \cdot g'(h^{-1} \circ h(c)) \cdot h^{-1'}(h(c)) = h'(g(c)) \cdot g'(c) \cdot h^{-1'}(h(c)) = 0$. □

As alluded to before, a less technical restatement of Corollary 7.10 from [**?**], included without proof, is:

Theorem 2.8 *Let $f(z) = z - z^{n+1} + a_{n+2}z^{n+2} + \cdots + a_d z^d$, where $n \geq 1$. Then, there exists a critical point, c, that converges to the origin under iteration of f. That is:*

$$\lim_{n \to \infty} f^n(c) = 0$$

A useful result that we will use later on is:

Lemma 2.9 *Let $a \in \mathbb{R}^+$. The image of the open disk $D(a, a)$ under $\frac{1}{z}$ is $\{z : Re(z) > \frac{1}{2a}\}$.*

Proof We will show that $z_0 \in D(a, a) = \{z = x + iy : (x - a)^2 + y^2 < a^2\} \Leftrightarrow \frac{1}{z_0} \in \{z : Re(z) > \frac{1}{2a}\}$.

Applying $\frac{1}{z}$ to $z_0 = x_0 + iy_0$, we have:

$$\frac{1}{z_0} = \frac{1}{x_0 + iy_0} = \frac{x_0 - iy_0}{x_0^2 + y_0^2} \Rightarrow Re\left(\frac{1}{z_0}\right) = \frac{x_0}{x_0^2 + y_0^2}.$$

Since $z_0 \neq 0$, the following chain of equivalences proves the result:

$$z_0 \in D(a,a) \Leftrightarrow x_0^2 - 2ax_0 + a^2 + y_0^2 = (x_0 - a)^2 + y_0^2 < a^2$$
$$\Leftrightarrow x_0^2 - 2ax_0 + y_0^2 < 0 \Leftrightarrow 2ax_0 > x_0^2 + y_0^2 \Leftrightarrow Re\left(\frac{1}{z_0}\right) = \frac{x_0}{x_0^2 + y_0^2} > \frac{1}{2a}.$$

□

3 Conjugacy and the SMVC

As a elementary result, we will first show that the bounds in Conjecture 1.1 and the SMVC are equivalent:

Lemma 3.1 *For all polynomials of the form $f(z) = z + a_2 z^2 + \cdots + a_d z^d$ there exists a critical point c of f for which*

$$\left|\frac{f(c)}{c}\right| \leq 1 \qquad \text{(the SMVC)}$$

$$\Longleftrightarrow$$

For all non-linear polynomials P where $P'(0) \neq 0$, and ζ not critical points of P, there exists a critical point c_i such that

$$\left|\frac{P(\zeta) - P(c_i)}{\zeta - c_i}\right| \leq |P'(\zeta)|. \qquad \text{(from (1))}$$

Proof The $\Leftarrow$ direction results from taking $f(z) = P(z)$ and $\zeta = 0$ in (1). To show the $\Rightarrow$ direction, define the functions

$$\beta(z) := z + \zeta \qquad \alpha_\nu(z) := \nu(z - P(\zeta)), \text{where } \nu \in \mathbb{C}$$
$$Q_\zeta(z) := \alpha_\nu \circ P \circ \beta(z) = \nu(P(z + \zeta) - P(\zeta)).$$

Note that $Q_\zeta(0) = 0$, and with the choice of $\nu = \frac{1}{P'(\zeta)}$, $Q'_\zeta(0) = 1$. Further, if c_i is a critical point of P, then $c_i - \zeta$ is a critical point of Q_ζ.

It follows that for every ζ not a critical point of P,

$$\left|\frac{Q_\zeta(c_i - \zeta)}{c_i - \zeta}\right| = \left|\frac{\nu(P(c_i) - P(\zeta))}{c_i - \zeta}\right| = \left|\frac{P(\zeta) - P(c_i)}{(\zeta - c_i)P'(\zeta)}\right|$$

Thus, for every non-linear polynomial P and ζ not a critical point of P, there exists a polynomial of the form $Q_\zeta(z) = z + a_2 z^2 + \cdots + a_d z^d$ where for some critical point c of Q_ζ and some critical point c_i of P:

$$\left|\frac{Q_\zeta(c)}{c}\right| = \left|\frac{P(\zeta) - P(c_i)}{(\zeta - c_i)P'(\zeta)}\right|$$

Thus,

$$\left|\frac{Q_\zeta(c)}{c}\right| \leq 1 \Rightarrow \left|\frac{P(\zeta) - P(c_i)}{\zeta - c_i}\right| \leq |P'(\zeta)|,$$

which completes the proof of the $\Rightarrow$ direction.

□

Theorem 3.1 can be viewed as a simplification of the SMVC: instead of looking at all non-linear polynomials, it is sufficient to prove the SMVC for polynomials of the form: $f(z) = z + a_2z^2 + \cdots + a_dz^d$. We will be able to simplify the SMVC and DSMVC further with conjugacy.

Finding the critical points of an arbitrary polynomial is difficult, and so we will instead parameterize polynomials by their critical points. Conjugacy allows us to consider only a specific parameterization. There are two preliminary lemmas before an important fact that we will often use.

Lemma 3.2 *Suppose f is conjugate to g by an invertible linear map, that is $f = h \circ g \circ h^{-1}$ for some h where $h(z) = \alpha z$, $\alpha \neq 0$. Then g has critical point c satisfying the SMVC $\Leftrightarrow$ f has critical point αc satisfying the SMVC.*

Proof We will prove the $\Rightarrow$ direction as $\Leftarrow$ is almost identical. g satisfies the SMVC $\Leftrightarrow \left|\frac{g(c)}{c}\right| \leq 1$, for some critical point c of g. From Lemma 2.7, αc is a critical point of f. It is easy to see that $\left|\frac{f(\alpha c)}{\alpha c}\right| = \left|\frac{\alpha g(\frac{\alpha c}{\alpha})}{\alpha c}\right| = \left|\frac{g(c)}{c}\right|$. So, $\left|\frac{g(c)}{c}\right| \leq 1$ implies $\left|\frac{f(\alpha c)}{\alpha c}\right| \leq 1$, which in turn implies that f satisfies the SMVC. □

Lemma 3.3 *Suppose f is conjugate to g by an invertible linear map, that is $f = h \circ g \circ h^{-1}$ for some h where $h(z) = \alpha z$, $\alpha \neq 0$. Then, g has a critical point c converging to the origin under iteration $\Leftrightarrow$ f has critical point αc converging to the origin under iteration. Further, g has a critical point $\tilde{c}$ that does not converge to the origin under iteration $\Leftrightarrow$ f has a critical point $\alpha\tilde{c}$ that does not converge to the origin under iteration.*

Proof We will prove the $\Rightarrow$ direction of both statements as $\Leftarrow$ is almost identical. Suppose c is a critical point of g such that $\lim_{n\to\infty} g^n(c) = 0$. Then, from Lemma 2.7, αc is a critical point for f. $\lim_{n\to\infty} f^n(\alpha c) = \lim_{n\to\infty} h \circ g^n \circ h^{-1}(\alpha c) = \lim_{n\to\infty} \alpha g^n(c) = 0$. Similarly, suppose $\tilde{c}$ is a critical point of g such that $\lim_{n\to\infty} g^n(\tilde{c}) \neq 0$. Then, from Lemma 2.7, $\alpha\tilde{c}$ is a critical point for f. $\lim_{n\to\infty} f^n(\alpha\tilde{c}) = \lim_{n\to\infty} h \circ g^n \circ h^{-1}(\alpha\tilde{c}) = \lim_{n\to\infty} \alpha g^n(\tilde{c}) \neq 0$, as $\alpha \neq 0$. □

So, when proving the DSMVC for a class of degree d polynomials, it is sufficient to prove the DSMVC for polynomials of the form:

$$f(z) = \int_0^z \prod_{i=1}^{d-1} (\zeta - c_i)\, d\zeta, \text{ where } (-1)^{d-1} \prod_{i=1}^{d-1} c_i = 1.$$

To see this, consider a degree d polynomial of the form

$$p(z) = z + a_2 z^2 + a_3 z^3 + \cdots + a_d z^d.$$

Note that $p'(z) = 1 + 2a_2 z + 3a_3 z^2 + \cdots + d a_d z^{d-1}$.

Now, conjugate p by $\alpha = (da_d)^{\frac{1}{d-1}}$, producing $f(z) := \alpha p\left(\frac{z}{\alpha}\right)$. Note that $f'(z) = \prod_{i=1}^{d-1}(z - c_i)$, where $(-1)^{d-1} \prod_{i=1}^{d-1} c_i = 1$. From Lemma 3.2, a critical point of p satisfies the SMVC $\Leftrightarrow$ the corresponding critical point of f satisfies the SMVC. From Lemma 3.3, a critical point of p converges to 0 under iteration $\Leftrightarrow$ the corresponding critical point of f converge to 0 under iteration. Using the Fundamental Theorem of Calculus, we may restrict our attention to proving the SMVC for polynomials of the form

$$f(z) = \int_0^z \prod_{i=1}^{d-1} (\zeta - c_i)\, d\zeta, \text{ where } (-1)^{d-1} \prod_{i=1}^{d-1} c_i = 1.$$

4 Minimal Critical Points

First, a definition:

Definition 4.1 Let $c_1, \ldots, c_{d-1}$ be the critical points of an degree d polynomial, f. c_j is a ***minimal critical point*** of f if $|c_j| = \min\{|c_i| : i = 1, \ldots, d-1\}$.

4.1 Minimal critical points satisfy the SMVC

A challenge of the SMVC is attempting to single out which critical points of a polynomial are likely to satisfy the bound. One simple choice is a critical point that is already closest to the origin (a minimal critical point). We will show that for quartic polynomials, this critical point does in fact satisfy the SMVC.

Theorem 4.2 *If f is a quartic polynomial of the form $f(z) = z + a_2 z^2 + a_3 z^3 + a_4 z^4$, then the minimal critical points of f satisfy the SMVC.*

Proof From Lemma 3.2, it is sufficient to prove the result for quartics of the form:

$$f(z) = \int_0^z (\zeta - c_1)(\zeta - c_2)(\zeta - c_3) d\zeta, \text{ where } -c_1 c_2 c_3 = 1.$$

Without loss of generality, suppose that c_1 is a minimal critical point. We will show that $\left|\frac{f(c_1)}{c_1}\right| \leq 1$:

Parameterizing $f(c_1)$ as a line integral by $c_1 t$, where $t \in [0,1]$, we have

$$f(c_1) = \int_0^{c_1} (\zeta - c_1)(\zeta - c_2)(\zeta - c_3) d\zeta = c_1 \int_0^1 (c_1 t - c_1)(c_1 t - c_2)(c_1 t - c_3) dt$$

It follows that

$$\left| \frac{f(c_1)}{c_1} \right| = \left| \int_0^1 (c_1 t - c_1)(c_1 t - c_2)(c_1 t - c_3) dt \right|.$$

It is an elementary fact that

$$\left| \int_0^1 (c_1 t - c_1)(c_1 t - c_2)(c_1 t - c_3) dt \right| \leq \int_0^1 \left| (c_1 t - c_1)(c_1 t - c_2)(c_1 t - c_3) \right| dt.$$

Using $c_1 c_2 c_3 = -1$,

$$\int_0^1 \left| (c_1 t - c_1)(c_1 t - c_2)(c_1 t - c_3) \right| dt = \int_0^1 \left| \frac{(c_1 t - c_1)(c_1 t - c_2)(c_1 t - c_3)}{c_1 c_2 c_3} \right| dt$$
$$= \int_0^1 \left| \frac{c_1 t - c_1}{c_1} \right| \left| \frac{c_1 t - c_2}{c_2} \right| \left| \frac{c_1 t - c_3}{c_3} \right| dt = \int_0^1 |t-1| \left| \frac{c_1}{c_2} t - 1 \right| \left| \frac{c_1}{c_3} t - 1 \right| dt.$$

From the triangle inequality,

$$\int_0^1 |t-1| \left| \frac{c_1}{c_2} t - 1 \right| \left| \frac{c_1}{c_3} t - 1 \right| dt \leq \int_0^1 |t-1| \left(\left| \frac{c_1}{c_2} \right| t + 1 \right) \left(\left| \frac{c_1}{c_3} \right| t + 1 \right) dt.$$

Using that $|c_1|$ is minimal,

$$\int_0^1 |t-1| \left(\left| \frac{c_1}{c_2} \right| t + 1 \right) \left(\left| \frac{c_1}{c_3} \right| t + 1 \right) dt \leq \int_0^1 |t-1| (t+1)(t+1) dt.$$

As $|t-1| = 1 - t$ for $t \in [0,1]$,

$$\int_0^1 |t-1| (t+1)(t+1) dt = \int_0^1 (1-t)(t+1)(t+1) dt = \frac{11}{12} < 1.$$

Thus, we have shown

$$\left| \frac{f(c)}{c} \right| \leq \ldots < 1.$$

□

Note that a similar argument cannot be extended for polynomials of higher degree: Using the same argument for a degree d polynomial, we obtain a final integral of the form:

$$\int_0^1 (1-t)(t+1)^{d-2}dt.$$

Using integration by parts, we see that

$$\begin{aligned}\int_0^1 (1-t)(t+1)^{d-2}\,dt &= \left.\frac{(1-t)(t+1)^{d-1}}{d-1}\right|_0^1 + \frac{1}{d-1}\int_0^1 (t+1)^{d-1}\,dt \\ &= -\frac{1}{d-1} + \left.\frac{(t+1)^d}{d(d-1)}\right|_0^1 \\ &= -\frac{1}{d-1} + \frac{2^d}{d(d-1)} - \frac{1}{d(d-1)} \\ &= \frac{2^d-d-1}{d(d-1)}.\end{aligned}$$

It is easy to check that for $d > 4$, $\dfrac{2^d-d-1}{d(d-1)} > 1$.

4.2 Minimal critical point doesn't necessarily converge

Since minimal critical points for quartics always satisfies the SMVC, it is natural to ask if minimal critical points always converges to the origin under iteration. If this were true, then the DSMVC would be resolved for quartics. However, the minimal critical point does not necessarily converge to the origin, as shown by the following counterexample:

Theorem 4.3 *Given $f(z) = z + a_2z^2 + a_3z^3 + a_4z^4$, a minimal critical point does not necessarily converge to the origin under iteration of f.*

Proof Let $c_1 = -\frac{4}{5}$, $c_2 = 1 + \frac{1}{2}i$, and $c_3 = 1 - \frac{1}{2}i$ so that

$$f(z) = \int_0^z (\zeta + \frac{4}{5})(\zeta - 1 - \frac{1}{2}i)(\zeta - 1 + \frac{1}{2}i)\,d\zeta = \frac{z^4}{4} - \frac{2}{5}z^3 - \frac{7}{40}z^2 + z$$

Note that $c_1 = -\dfrac{4}{5}$ is the minimal critical point. Let $p := \dfrac{8-\sqrt{134}}{10} \approx -0.358$. It is easy to check that p is a fixed point of f and that $p \in \left[-\frac{4}{5}, -\frac{3}{10}\right]$. We will now show $\lim\limits_{n\to\infty} f^n(c_1) = p \neq 0$. Note, $f^n(-\frac{4}{5}) \in \mathbb{R}$, $\forall n \in \mathbb{N}$. So, for our purposes, we can restrict $f|_{\mathbb{R}} : \mathbb{R} \to \mathbb{R}$,

$$f(x) = \frac{x^4}{4} - \frac{2}{5}x^3 - \frac{7}{40}x^2 + x$$

It is an easy calculus exercise to show that f' is increasing on $(-\infty, -\frac{3}{10}]$. Now observe that

$$f'(-\frac{4}{5}) = 0 \text{ and } f'(-\frac{3}{10}) = \frac{97}{100}$$

Since f' is increasing on $[-\frac{4}{5}, -\frac{3}{10}]$, $0 < f'(x) < \frac{98}{100} < 1 \ \forall x \in (-\frac{4}{5}, -\frac{3}{10})$. Since f is C^1, we can apply the Mean Value Theorem to obtain:

$$0 < p - f(-\frac{4}{5}) = f(p) - f(-\frac{4}{5}) = f'(y_1)(p - (-\frac{4}{5})) \qquad \text{for some } y_1 \in \left(-\frac{4}{5}, p\right).$$

Further, using that $0 < f'(x) < 0.98$ on $\left(-\frac{4}{5}, p\right)$, we see that

$$0 < p - f(-\frac{4}{5}) < 0.98\left(p + \frac{4}{5}\right).$$

Notice that this implies that $f(-\frac{4}{5}) \in (-\frac{4}{5}, p]$. We will now repeat the same process to obtain:

$$0 < f^2(p) - f^2(-\frac{4}{5}) = f'(y_2)(f(p) - f(-\frac{4}{5})) \quad \text{for some } y_2 \in (f(-\frac{4}{5}), p)$$

$$\Rightarrow 0 < p - f^2(-\frac{4}{5}) = f'(y_2)(p - f(-\frac{4}{5}))$$

$$\Rightarrow 0 < p - f^2(-\frac{4}{5}) < 0.98\left(p - f(-\frac{4}{5})\right) < (0.98)^2\left(p - (-\frac{4}{5})\right).$$

This implies that $f^2(-\frac{4}{5}) \in (f(-\frac{4}{5}), p] \subsetneq (-\frac{4}{5}, p]$. Continuing in this way, we obtain

$$0 < p - f^n(-\frac{4}{5}) < 0.98^n\left(p - (-\frac{4}{5})\right)$$

Using the Squeeze Theorem and $\lim_{n\to\infty}(0.98)^n = 0$, it is evident that

$$\lim_{n\to\infty} f^n(-\frac{4}{5}) = p = \frac{8 - \sqrt{134}}{10} \neq 0$$

Thus we have shown that the minimal critical point of f does not necessarily converge to the origin under iteration of f. A picture of this is shown in Figure 1.

□

Figure 1: The blue line is graph of $f(x)$, the red line is graph of identity function. Note that the minimal critical point is $-\frac{4}{5}$, which converges to fixed point $p \neq 0$.

5 Quartics with Repeated Critical Points

Our first main result is a proof of the DSMVC for quartics with repeated critical points. The motivation for considering this special case is that it is a reduction of the parameter space of quartic polynomials (parameterized by critical points) from a 2 complex dimensional space to a 1 complex dimensional space. Additionally, during the proof we work with the modulus of the critical point, which reduces the dimension further, to a 1-dimensional real space. In the proof, we will demonstrate that under certain conditions, a minimal critical point actually does converge to the origin under iteration.

Theorem 5.1 (DSMVC for Quartics with Repeated Critical Points) *Let $p(z) = z + a_2z^2 + a_3z^3 + a_4z^4$ be a quartic polynomial. Suppose there are critical points of p with multiplicity greater than 1. Then, p satisfies the DSMVC.*

Proof We may restrict our attention to the quartic polynomials of the form

$$f(z) = \int_0^z (\zeta - c_1)(\zeta - c_2)(\zeta - c_3)d\zeta, \text{ where } -c_1c_2c_3 = 1.$$

Now, we will examine the case where there is a repeated critical point of f. We have two cases:

1. There is one distinct critical point
2. There are two distinct critical points

The 1^{st} case is trivial as we know that there always is a critical point converging to the origin under iteration, from Theorem 2.8. We will focus on proving the 2^{nd} case.

5.1 Case of 2 Distinct Critical Points

We can rewrite f as

$$f(z) = \int_0^z (\zeta - c)^2(\zeta + \frac{1}{c^2})\, d\zeta = z - \left(\frac{1}{c} - \frac{c^2}{2}\right) z^2 + \left(\frac{1}{3c^2} - \frac{2c}{3}\right) z^3 + \frac{z^4}{4}, \text{ where } c \neq 0$$

When written this way, the critical points of f are c and $-\frac{1}{c^2}$. Algebra shows that

$$c \text{ satisfies the SMVC} \Leftrightarrow \left|4 + c^3\right| \leq 12.$$

$$-\frac{1}{c^2} \text{ satisfies the SMVC} \Leftrightarrow \left|-6 - \frac{4}{c^3} - \frac{1}{c^6}\right| \leq 12.$$

In particular, when applying the triangle inequality, it is evident that sufficient conditions are:

$$4 + |c|^3 \leq 12 (\Leftrightarrow |c| \leq 2) \Rightarrow c \text{ satisfies the SMVC.}$$

$$\frac{4}{|c|^3} + \frac{1}{|c|^6} \leq 6 \left(\Leftrightarrow c \geq \sqrt[3]{\frac{2 + \sqrt{10}}{6}} \approx 0.95111\right) \Rightarrow -\frac{1}{c^2} \text{ satisfies the SMVC.}$$

The regions where c and $-\frac{1}{c^2}$ separately satisfy the SMVC in Figures 2 and 3, respectively.

Rounding, the region $0.952 \leq |c| \leq 2$ has both c and $-\frac{1}{c^2}$ satisfy the SMVC, shown in Figure 4. We know at least 1 critical point converges to the origin under iteration. Thus, Theorem 5.1 holds when $0.952 \leq |c| \leq 2$.

We will need to consider the remaining cases:

1. $|c| > 2$
2. $|c| < 0.952$

First, we conjugate g even further, by $\beta = \frac{1}{c} - \frac{c^2}{2}$ to produce

$$h(z) := \beta g(\frac{z}{\beta}) = z - z^2 + Az^3 - Bz^4 \tag{2}$$

where

$$A = \frac{4 - 8c^3}{3(-2 + c^3)^2} \qquad B = \frac{2c^3}{(-2 + c^3)^3} \tag{3}$$

The exceptional case where $\beta = 0 \Leftrightarrow c = \sqrt[3]{2}$, which lies in the region where both critical points satisfy the SMVC. We can apply Theorem 2.8 to see that at least one critical point always converges to the origin under iteration of f. Thus, our conjugation is appropriate to deal with the cases $|c| > 2$ and $|c| < 0.952$.

Figure 2: The orange region is defined by $|c| \leq 2$ and the purple region is defined by $|c| > 2$ and $\left|4 + c^3\right| \leq 12$.

Figure 3: The orange region is defined by $c \geq \sqrt[3]{\frac{2+\sqrt{10}}{6}}$ and the purple region is defined by $|c| < \sqrt[3]{\frac{2+\sqrt{10}}{6}}$ and $\left|-6 - \frac{4}{c^3} - \frac{1}{c^6}\right| \leq 12$.

Figure 4: The orange region is defined by: $0.952 \leq |c| \leq 2$; the purple region is defined by the union of regions: 1. $|c| < \sqrt[3]{\frac{2+\sqrt{10}}{6}}$ and $\left|-6 - \frac{4}{c^3} - \frac{1}{c^6}\right| \leq 12$ and 2. $|c| > 2$ and $\left|4 + c^3\right| \leq 12$.

5.2 $|c| > 2$

When $|c| > 2$, $-\frac{1}{c^2}$ satisfies the SMVC. Note $-\frac{1}{c^2}$ is the minimal critical point. We will show that $-\frac{1}{c^2}$ also converges to the origin under iteration.

From Lemmas 2.7 and 3.3, the conjugated map h has critical point $v := \beta\left(-\frac{1}{c^2}\right) = \frac{1}{2} - \frac{1}{c^3}$, which converges to the origin under iteration of h exactly when $-\frac{1}{c^2}$ converges to the origin under iteration of f.

Let v_2 denote the 2^{nd} iterate of v under h. Computation shows that

$$\begin{aligned} v_2 = \frac{3}{16} &+ \frac{1}{82944c^{33}} + \frac{31}{165888c^{30}} + \frac{1}{864c^{27}} + \frac{83}{20736c^{24}} + \frac{95}{10368c^{21}} \\ &+ \frac{83}{6912c^{18}} - \frac{13}{2592c^{15}} - \frac{149}{2592c^{12}} - \frac{29}{192c^9} - \frac{281}{1152c^6} - \frac{5}{24c^3} \end{aligned}$$

To bound the location of v_2, we use the triangle inequality to obtain:

$$\begin{aligned} \left|\frac{3}{16} - v_2\right| &\leq \left|\frac{1}{82944c^{33}}\right| + \left|\frac{31}{165888c^{30}}\right| + \left|\frac{1}{864c^{27}}\right| + \left|\frac{83}{20736c^{24}}\right| + \left|\frac{95}{10368c^{21}}\right| \\ &\quad + \left|\frac{83}{6912c^{18}}\right| + \left|\frac{13}{2592c^{15}}\right| + \left|\frac{149}{2592c^{12}}\right| + \left|\frac{29}{192c^9}\right| + \left|\frac{281}{1152c^6}\right| + \left|\frac{5}{24c^3}\right| \\ &< 0.0302, \end{aligned}$$

where we have used $|c| > 2$.

So, $v_2 \in D(\frac{3}{16}, 0.0302) \subset D(0.125, 0.125) \subset D(0, 0.25)$. It follows that $|v_2| < 0.25$ and from Lemma 2.9, $Re\left(\frac{1}{v_2}\right) > 4$. Next, we conjugate h (from (2)) by $\frac{1}{z} = w$ to produce

$$y(w) := \frac{1}{h\left(\frac{1}{w}\right)} = w + 1 + \frac{1-A}{w} + \frac{B-2A+1}{w^2} + \frac{C}{w^3}, \tag{4}$$

where

$$C = \frac{(2B + A^2 - 3A + 1) + (B - A(2B - 2A + 1))z + (B(B - 2A + 1))z^2}{1 - z + Az^2 - Bz^3}. \tag{5}$$

Using the triangle inequality and $|c| > 2$, we can bound $|A|$, $|B|$, and $|C|$ as follows:

$$|A| = \left|\frac{4 - 8c^3}{3(-2 + c^3)^2}\right| \leq \frac{4 + 8\,|c|^3}{3(|c|^3 - 2)^2} < \frac{4 + 8 \cdot 2^3}{3(2^3 - 2)^2} = \frac{17}{27}.$$

The last inequality uses that $\frac{4+8|c|^3}{3(|c|^3-2)^2}$ is decreasing with respect to $|c|$ when $|c| > 2$. Observe,

$$|B| = \left|\frac{2c^3}{(-2 + c^3)^3}\right| \leq \frac{2\,|c|^3}{(|c|^3 - 2)^3} < \frac{2 \cdot 2^3}{(2^3 - 2)^3} = \frac{2}{27}.$$

The last inequality uses that $\frac{2|c|^3}{(|c|^3-2)^3}$ is decreasing with respect to $|c|$ when $|c| > 2$.

We can bound $|C|$ in a similar fashion:

$$|C| = \left|\frac{(2B + A^2 - 3A + 1) + (B - A(2B - 2A + 1))z + (B(B - 2A + 1))}{1 - z + Az^2 - Bz^3}\right|$$

$$\leq \frac{(2\,|B| + |A|^2 + 3\,|A| + 1) + (|B| + 2\,|AB| + 2\,|A|^2 + |A|)\,|z| + (|B|^2 + 2\,|AB| + |B|)\,|z|^2}{1 - |z| - A\,|z|^2 - B\,|z|^3}. \tag{6}$$

Using the bounds for $|A|$ and $|B|$, we can further bound $|C|$ when $z \in D(0, 0.25)$:

$$|C| \leq \cdots \leq \frac{\frac{2503}{729} + \frac{1159}{729}\,|z| + \frac{14}{81}\,|z|^2}{1 - |z| - \frac{17}{27}\,|z|^2 - \frac{2}{27}\,|z|^3} < 5.42.$$

Let $w_2 = \dfrac{1}{v_2}$. Now, we will show that $Re(w_2)$ converges to $+\infty$ under iteration of y, which equivalently shows that v_2 converges to 0 under iteration of g. We know that $|v_2| < 0.25 (\Leftrightarrow |w_2| > 4)$ and $Re(w_2) > 4$, so the following inequalities hold under induction:

$$\begin{aligned} Re(y(w)) &\geq Re(w) + 1 - \frac{1 + |A|}{|w|} - \frac{|B| + 2|A| + 1}{|w|^2} - \frac{|C|}{|w|^3} \\ &\geq Re(w) + 1 - \frac{44}{27} \cdot 0.25 - \frac{7}{3} \cdot 0.25^2 - 5.42 \cdot 0.25^3 \\ &> Re(w) + 1 - 0.64 = Re(w) + 0.36. \end{aligned}$$

Thus, the right half-plane $Re(w) \geq 4$ is forward-invariant under $y(w)$. From induction, w_2 converges to ∞ under iteration of y, implying v_2 converges to 0 under iteration of h.

This completes the proof of the case $|c| > 2$.

5.3 $|c| < 0.952$

When $|c| < 0.952$, c satisfies the SMVC. We will show that c also converges to the origin under iteration.

From Lemmas 2.7 and 3.3, the conjugated map has critical point $v := \beta(c) = 1 - \dfrac{c^3}{2}$, which converges to the origin under iteration of h exactly when c converges to the origin under iteration of f.

Let v_5 denote the 5^{th} iterate of v under h. Computation shows that

$$v_5 = b_0 + b_3 c^3 + b_6 c^6 + \cdots + b_{1026} c^{1026}$$

where $b_0 = \dfrac{7016552399014818317342795082052043500965788887821737490 97}{5391030899743293631239539488528815119194426882613553319203}$ and $b_3, b_6, \ldots, b_{1026}$ are not shown for conciseness.

To bound the location of v_5, we use the triangle inequality to obtain:

$$|b_0 - v_5| \leq \left|b_3 c^3\right| + \left|b_6 c^6\right| + \cdots + \left|b_{1026} c^{1026}\right| < \frac{27}{1000},$$

where the last inequality used $|c| < \frac{952}{1000}$ and computer computation using fractions to find an upper bound. Using $b_0 < 0.131$, we have $v_5 \in D(b_0, 0.027) \subset D(0.79, 0.79) \subset D(0, 0.158)$ It follows that $|v_5| < 0.158$ and from Lemma 2.9, $Re\left(\dfrac{1}{v_5}\right) > \dfrac{1}{0.158}$.

We will now use $y(w)$ from (4), with A, B from (3) and C from (5).

Using the triangle inequality and $|c| < \frac{952}{1000}$, we can bound $|A|$, $|B|$, and $|C|$ using computer calculations involving fractions. We obtain the following inequalities:

$$|A| = \left|\frac{4 - 8c^3}{3(-2 + c^3)^2}\right| \leq \frac{4 + 8|c|^3}{3(2 - |c|^3)^2} < \frac{4 + 8 \cdot \frac{952}{1000}^3}{3(2 - \frac{952}{1000}^3)^2} < \frac{282}{100}.$$

The last inequality uses that $\frac{4+8|c|^3}{3(2-|c|^3)^2}$ is increasing with respect to $|c|$ when $|c| < \frac{952}{1000}$. Observe,

$$|B| = \left|\frac{2c^3}{(-2+c^3)^3}\right| \leq \frac{2\,|c|^3}{(2-|c|^3)^3} < \frac{2\cdot\frac{952}{1000}^3}{(2-\frac{952}{1000}^3)^3} < \frac{118}{100}.$$

The last inequality uses that $\frac{2|c|^3}{(2-|c|^3)^3}$ is increasing with respect to $|c|$ when $|c| < \frac{952}{1000}$.

Bounding $|C|$ as in (6) and using the bounds for $|A|$ and $|B|$, we can further bound $|C|$ when $z \in D(0, 0.158)$ by:

$$|C| \leq \cdots \leq \frac{19.78 + 26.56\,|z| + 9.23\,|z|^2}{1 - |z| - 2.82\,|z|^2 - 1.18\,|z|^3} < 31.6.$$

Let $w_5 = \dfrac{1}{v_5}$. Now, we will show that $Re(w_5)$ converges to $+\infty$ under iteration of y, which equivalently shows that v_5 converges to 0 under iteration of g. We know that $|v_5| < 0.158 (\Leftrightarrow |w_2| > \frac{1}{0.158})$ and $Re(w_2) > \frac{1}{0.158}$, so the following inequalities hold under induction:

$$\begin{aligned} Re(y(w)) &\geq Re(w) + 1 - \frac{1+|A|}{|w|} - \frac{|B| + 2\,|A| + 1}{|w|^2} - \frac{|C|}{|w|^3} \\ &\geq Re(w) + 1 - 3.82 \cdot 0.158 - 7.82 \cdot 0.158^2 - 31.6 \cdot 0.158^3 \\ &> Re(w) + 1 - 0.93 = Re(w) + 0.07. \end{aligned}$$

Thus, the right half-plane $Re(w) \geq \frac{1}{0.158}$ is forward-invariant under $y(w)$. From induction, w_5 converges to ∞ under iteration of y, implying v_5 converges to 0 under iteration of h.

This completes the proof of the case $|c| < 0.952$ and the proof of the theorem. □

6 Quartics with Real Critical Points

Our second main result is a proof of the DSMVC for quartics with real critical points. The motivation for considering this special case is that by restricting the critical points to be real, elementary methods of single-variable calculus could be applied to solve the problem without using computer calculations.

We can conjugate and restrict our attention to quartics of the form

$$f(z) = \int_0^z (\zeta - c_1)(\zeta - c_2)(\zeta + \frac{1}{c_1 c_2})d\zeta = z + \frac{\left(-c_1 - c_2 + c_1^2 c_2^2\right) z^2}{2c_1 c_2} - \frac{\left(-1 + c_1 c_2 + c_1 c_2^2\right) z^3}{3c_1 c_2} + \frac{z^4}{4}.$$

Define

$$h(z) := \frac{f(z) - z}{z^2} = \frac{\left(-c_1 - c_2 + c_1^2 c_2^2\right)}{2c_1 c_2} - \frac{\left(-1 + c_1 c_2 + c_1 c_2^2\right) z}{3c_1 c_2} + \frac{z^2}{4}.$$

To prove DSMVC, we will show DSMVC holds for quartics for which:

1. $h(0) = 0$ (Lemma 6.1)
2. $h(0) < 0$ (Lemma 6.2)
3. $h(0) > 0$ (Lemma 6.3)

Lemma 6.1 *If $h(0) = 0$, then the minimal positive and negative critical points satisfy the DSMVC.*

Proof First, we need to show that there exists positive and negative critical points. From hypothesis,

$$h(0) = \frac{-c_1 - c_2 + c_1^2 c_2^2}{2c_1 c_2} = 0.$$

It is easy to see that both c_1 and c_2 cannot be less than 0. If they were, then $h(0) > 0$, a contradiction. If both c_1 and c_2 are not less than 0, then at least one of: c_1, c_2, and $-\frac{1}{c_1 c_2}$ is positive, and at least one critical point is negative.

Let $\tilde{c_1}$ be the minimal positive critical point and $\tilde{c_2}$ be the minimal negative critical point. Note that $\tilde{c_1}$ and $\tilde{c_2}$ are not necessarily c_1 and c_2, respectively. We claim that $0 < f(z) < z$ on $(0, \tilde{c_1}]$ and $0 > f(z) > z$ on $[\tilde{c_2}, 0)$, which proves the DSMVC. We will prove the case $0 < f(z) < z$ on $(0, \tilde{c_1}]$, then briefly describe the proof for $0 > f(z) > z$ on $[\tilde{c_2}, 0)$:

It follows easily that the whole interval $(0, \tilde{c_1}]$ satisfies $f(z) > 0$: Suppose if there was a $\theta \in (0, \tilde{c_1}]$ where $f(\theta) \leq 0$. From $f'(0) = 1$, we know locally around 0, f is increasing and behaves like the identity function. So, from the Intermediate Value Theorem, there would exist a $j \in (0, \theta]$ where $f(j) = 0$. However, from Rolle's Theorem this would imply there would exist a $\tilde{c} \in (0, j)$ where $f'(\tilde{c}) = 0$, contradicting that $\tilde{c_1}$ was the minimal positive critical point. Thus, the whole interval $(0, \tilde{c_1}]$ satisfies $f(z) > 0$.

To prove that $(0, \tilde{c_1}]$ satisfies $f(z) < z$, it suffices to show that $f'(z) < 1$ on $(0, \tilde{c_1})$: if $\exists q \in [0, \tilde{c_1}]$ where $f(q) > q$, then by the Mean Value Theorem there would be a $\tilde{q} \in (0, q)$ where

$$f'(\tilde{q}) = \frac{f(q) - f(0)}{q - 0} = \frac{f(q)}{q} > 1,$$

a contradiction.

To show $f'(z) < 1$ on $(0, \tilde{c_1})$, we will show f' has its local maximum at $0 \notin (0, \tilde{c_1})$. From hypothesis,

$$h(0) = \frac{-c_1 - c_2 + c_1^2 c_2^2}{2c_1 c_2} = 0.$$

Solving for c_2 in terms of c_1, we find that:

$$h(0) = 0 \Leftrightarrow c_2 = \frac{1 - \sqrt{1 + 4c_1^3}}{2c_1^2} \ , \ \frac{1 + \sqrt{1 + 4c_1^3}}{2c_1^2} \ , \text{ or } 0.$$

Clearly $c_2 \neq 0$, so we only need to consider the two cases

$$c_2 = \frac{1 - \sqrt{1 + 4c_1^3}}{2c_1^2} \text{ or } \frac{1 + \sqrt{1 + 4c_1^3}}{2c_1^2}.$$

For both cases, f defined by c_1 and c_2 is:

$$f(z) = z - \frac{\left(1 + c_1^3\right) z^3}{3c_1^2} + \frac{z^4}{4}.$$

From seeing that $f''(0) = 0$ and $f'''(0) < 0$, we find that 0 is the local maximum for $f'(z)$.

Using these facts with the limit definition of the derivative and the Mean Value Theorem, we know $\exists \delta > 0$ where $z \in (0, \delta)$ satisfies $0 < f'(z) < f'(0) = 1$. Further, since we know that f' has no local maximum on $(0, \tilde{c_1})$, we know that the maximum of f' on $[\frac{\delta}{2}, \tilde{c_1}]$ is at one of the endpoints. Since $f'(\tilde{c_1}) = 0$, the maximum is at $\frac{\delta}{2}$. Thus, we have shown that $[\frac{\delta}{2}, \tilde{c_1}]$ and $(0, \delta)$ satisfy $f'(z) < 1$, which shows $f'(z) < 1$ on $(0, \tilde{c_1})$. This completes the proof that the minimal positive critical point satisfies the DSMVC.

To prove that the minimal negative critical point satisfies the DSMVC, we use the same approach: we show that $[\tilde{c_2}, 0)$ satisfies $0 < f(z) < z$. From Intermediate Value Theorem and Rolle's Theorem, $f(z) < 0$ on $[\tilde{c_2}, 0)$. Then, using that f' takes its maximum at 0 and $f'''(0) < 0$, we can show that $f'(z) < 1$ on $[\tilde{c_2}, 0)$. Finally, using a contradiction with the Mean Value Theorem, this shows that $f(z) > z$ on $[\tilde{c_2}, 0)$.

□

Lemma 6.2 *If $h(0) < 0$, then the minimal positive critical point satisfies the DSMVC.*

Proof First, we need to show there exists a positive critical point. From hypothesis,

$$h(0) = \frac{-c_1 - c_2 + c_1^2 c_2^2}{2c_1 c_2} < 0.$$

It is easy to see that either c_1 or c_2 must be greater than 0: if they both were less than zero, then we would have $h(0) > 0$.

Let $\tilde{c_1}$ be the minimal positive critical point. As in the proof of Lemma 6.1, it follows easily that $f(z) > 0$ on $(0, \tilde{c_1}]$. What remains to be shown it that $f(z) < z$. Let v be the local maximum of f'. We will first show that $v \notin (0, \tilde{c_1}]$.

From hypothesis,

$$h(0) = \frac{-c_1 - c_2 + c_1^2 c_2^2}{2c_1 c_2} < 0,$$

Define the critical points of f' as v_1 and v_2:

$$v_1 := \frac{-1 + c_1^2 c_2 + c_1 c_2^2 - \sqrt{1 + c_1^2 c_2 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 c_2^4}}{3c_1 c_2} \tag{7}$$

$$v_2 := \frac{-1 + c_1^2 c_2 + c_1 c_2^2 + \sqrt{1 + c_1^2 c_2 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 c_2^4}}{3c_1 c_2}. \tag{8}$$

Using the second derivative test on the critical points of f', we obtain:

$$f'''(v_1) = -\frac{2\sqrt{1 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 \left(c_2 + c_2^4\right)}}{c_1 c_2} \tag{9}$$

$$f'''(v_2) = \frac{2\sqrt{1 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 \left(c_2 + c_2^4\right)}}{c_1 c_2}. \tag{10}$$

Now, there are 2 cases to consider:

1. $c_1 c_2 > 0$
2. $c_1 c_2 < 0$

In both cases, we will show $h(0) < 0 \Rightarrow v < 0 \Rightarrow v \notin (0, \tilde{c_1}]$.

Case 1 $\mathbf{c_1 c_2 > 0}$: From the results of the second derivative test in (9), v_1 is the local maximum for f'. So, $v = v_1$. Using (7), it is easy to see that:

$$v_1 < 0 \Leftrightarrow -1 + c_1^2 c_2 + c_1 c_2^2 < \sqrt{1 + c_1^2 c_2 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 c_2^4}.$$

Thus, to show $v_1 < 0$, is sufficient to show:

$$\left(-1 + c_1^2 c_2 + c_1 c_2^2\right)^2 < 1 + c_1^2 c_2 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 c_2^4.$$

However, this follows from $h(0) < 0$, as evidenced by the following:

$$\left(-1 + c_1^2 c_2 + c_1 c_2^2\right)^2 < 1 + c_1^2 c_2 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 c_2^4 \Leftrightarrow -c_1 - c_2 + c_1^2 c_2^2 < 0 \Leftrightarrow h(0) < 0.$$

Case 2 $\mathbf{c_1 c_2 < 0}$: From the results of the second derivative test in (10), v_2 is the local maximum for f'. So, $v = v_2$. Using (8), it is easy to see that:

$$v_2 < 0 \Leftrightarrow 1 - c_1^2 c_2 - c_1 c_2^2 < \sqrt{1 + c_1^2 c_2 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 c_2^4}.$$

Thus, to show $v_2 < 0$, it is sufficient to show:

$$\left(1 - c_1^2 c_2 - c_1 c_2^2\right)^2 < 1 + c_1^2 c_2 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 c_2^4.$$

However, this follows from $h(0) < 0$, as evidenced by the following:

$$\left(1 - c_1^2 c_2 - c_1 c_2^2\right)^2 < 1 + c_1^2 c_2 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 c_2^4 \Leftrightarrow c_1 + c_2 - c_1^2 c_2^2 < 0 \Leftrightarrow h(0) < 0.$$

To verify $f(z) < z$ on $(0, \tilde{c_1}]$, it is sufficient to use $h(0) < 0$ and that $f'(z)$ doesn't have a local maximum on $(0, \tilde{c_1})$: Although 0 is a fixed point for f, $h(0) < 0$ implies that in a neighborhood of 0, $f(z) < z$, except at 0 when $f(0) = 0$. Consider the positive part of this neighborhood, $(0, \delta]$ for $\delta > 0$ sufficiently small. Now, from the Mean Value Theorem, we know $\exists q \in (0, \delta)$ such that

$$f'(q) = \frac{f(\delta) - f(0)}{\delta - 0} = \frac{f(\delta)}{\delta} < 1.$$

Now, consider $[q, \tilde{c_1}]$. We know that $0 < f'(q) < 1$, and $f'(\tilde{c_1}) = 0 < 1$. If f' has no local maximum on $(0, \tilde{c_1})$, then f' has no local maxima on $(q, \tilde{c_1})$. Thus, the maximum of f' on $[q, \tilde{c_1}]$ is at q, so $f'(z) \leq f'(q) < 1$ on $[q, \tilde{c_1}]$.

Next, we will show that $f'(z) < 1$ on $[q, \tilde{c_1}]$ implies that $f(z) < z$ on $[q, \tilde{c_1}]$: Consider if not, and $\exists w \in [q, \tilde{c_1}]$ such that $f(w) > w$. Then, from the Mean Value Theorem, $\exists \tilde{q} \in (q, c)$ where

$$f'(\tilde{q}) = \frac{f(w) - f(q)}{w - q} > \frac{w - q}{w - q} = 1.$$

However, this contradicts $f'(z) < 1$ on $[q, \tilde{c_1}]$.

We have shown $f(z) < z$ on $[q, \tilde{c_1}]$ and $(0, \delta)$, which implies $f(z) < z$ on $(0, \tilde{c_1}]$. Previously, we showed $0 < f(z)$ on $(0, \tilde{c_1}]$. This completes the proof. □

Lemma 6.3 *If $h(0) > 0$, then the minimal negative critical point satisfies the DSMVC.*

Proof It is easy to see that there always exists a negative critical point. Let $\tilde{c_2}$ be the minimal negative critical point. As in the proof of Lemma 0.2, it follows easily that $f(z) < 0$ on $[\tilde{c_2}, 0)$. What remains to be shown is that $f(z) > z$. Let v be the local maximum of f'. We will first show that $v \notin (\tilde{c_2}, 0)$. From hypothesis,

$$h(0) = \frac{-c_1 - c_2 + c_1^2 c_2^2}{2c_1 c_2} > 0.$$

The critical points for f' are v_1 and v_2 as defined in (7) and (8), respectively. The results of the second derivative tests for v_1 and v_2 are the same as in (9) and (10), respectively.

Now, there are 2 cases to consider:

1. $c_1 c_2 > 0$
2. $c_1 c_2 < 0$

In both cases, we will show $h(0) > 0 \Rightarrow v > 0 \Rightarrow v \notin [\tilde{c_2}, 0)$.

Case 1 $\mathbf{c_1 c_2 > 0}$: Note that v_1 is the local maximum for f'. So, $v = v_1$. Using (7), it is easy to see that

$$v_1 > 0 \Leftrightarrow -1 + c_1^2 c_2 + c_1 c_2^2 > \sqrt{1 + c_1^2 c_2 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 c_2^4}.$$

Thus, to show that $v_1 > 0$, it is sufficient to show that

$$\left(-1 + c_1^2 c_2 + c_1 c_2^2\right)^2 > 1 + c_1^2 c_2 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 c_2^4.$$

However, this follows from $h(0) > 0$, as evidenced by the following:

$$\left(-1 + c_1^2 c_2 + c_1 c_2^2\right)^2 > 1 + c_1^2 c_2 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 c_2^4 \Leftrightarrow -c_1 - c_2 + c_1^2 c_2^2 > 0 \Leftrightarrow h(0) > 0.$$

Case 2 $\mathbf{c_1 c_2 < 0}$: Note that v_2 is the local maximum for f'. So, $v = v_2$. It is easy to see from (8) that

$$v_2 > 0 \Leftrightarrow 1 - c_1^2 c_2 - c_1 c_2^2 > \sqrt{1 + c_1^2 c_2 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 c_2^4}.$$

Thus, to show $v_2 > 0$, it is sufficient to show

$$\left(1 - c_1^2 c_2 - c_1 c_2^2\right)^2 > 1 + c_1^2 c_2 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 c_2^4.$$

However, this follows from $h(0) > 0$, as evidenced by the following:

$$\left(1 - c_1^2 c_2 - c_1 c_2^2\right)^2 > 1 + c_1^2 c_2 + c_1 c_2^2 + c_1^4 c_2^2 - c_1^3 c_2^3 + c_1^2 c_2^4 \Leftrightarrow c_1 + c_2 - c_1^2 c_2^2 > 0 \Leftrightarrow h(0) > 0.$$

To verify $f(z) > z$ on $[\tilde{c_2}, 0)$, it is sufficient to use $h(0) > 0$ and that $f'(z)$ doesn't have a local maximum on $(\tilde{c_2}, 0)$: Although 0 is a fixed point for f, $h(0) > 0$ implies that in a neighborhood of 0, $f(z) > z$, except at 0 when $f(0) = 0$. Consider the negative part of this neighborhood, $[-\delta, 0)$ for $\delta > 0$

sufficiently small. Now, from the Mean Value Theorem, we know $\exists q \in (-\delta, 0)$ such that

$$f'(q) = \frac{f(\delta) - f(0)}{\delta - 0} = \frac{f(\delta)}{\delta} < 1.$$

Now, consider $[\tilde{c_2}, q]$. We know that $0 < f'(q) < 1$, and $f'(\tilde{c_2}) = 0 < 1$. If f' has no local maximum on $(\tilde{c_2}, 0)$, then f' has no local maxima on $(\tilde{c_2}, q)$. Thus, the maximum of f' on $[\tilde{c_2}, q]$ is at q, so $f'(z) \leq f'(q) < 1$ on $[\tilde{c_2}, q]$.

Next, we will show that $f'(z) < 1$ on $[\tilde{c_2}, q]$ implies that $f(z) > z$ on $[\tilde{c_2}, q]$: Consider if not, and $\exists w \in [\tilde{c_2}, q]$ such that $f(w) < w$. Then, from the Mean Value Theorem, $\exists \tilde{q} \in (\tilde{c_2}, q)$ where

$$f'(\tilde{q}) = \frac{f(q) - f(w)}{q - w} > \frac{q - w}{q - w} = 1.$$

However, this contradicts $f'(z) < 1$ on $[\tilde{c_2}, q]$.

We have shown $f(z) > z$ on $[\tilde{c_2}, q]$ and $[-\delta, 0)$, which implies $f(z) > z$ on $[\tilde{c_2}, 0)$. Previously, we showed $f(z) < 0$ on $[\tilde{c_2}, 0)$. This completes the proof. □

7 General Quartics

After focusing on the two special cases of quartics with repeated critical points and quartics with real critical points, we look at general quartics. We were only able to obtain partial results for the general case. The main difficulty is that the minimal critical point does not always converge (see Theorem 4.3). In dealing with the case of repeated critical points, we were able to show in some settings, the minimal critical point did in fact converge. In dealing with the case of the real critical points, we were able to consider minimal positive and negative critical points. However, in the case of the general quartic, there are some regions in $|c_1|$, $|c_2|$ parameter space where the minimal critical point sometimes converges and sometimes doesn't, for arbitrary (c_1, c_2). Using MatLab, we were able to plot where these regions were, shown in Figure 5.

For the remainder of this section, we will describe our partial results for quartics in general. As before, we can consider quartic polynomials of the form

$$\begin{aligned} f(z) &= \int_0^z (\zeta - c_1)(\zeta - c_2)(\zeta + \frac{1}{c_1 c_2}) d\zeta \\ &= z + \frac{1}{2}\left(-\frac{1}{c_1} - \frac{1}{c_2} + c_1 c_2\right) z^2 - \frac{1}{3}\left(c_1 - \frac{1}{c_1 c_2} + c_2\right) z^3 + \frac{z^4}{4}. \end{aligned} \tag{11}$$

We will spend the rest of the report proving the following theorem:

Theorem 7.1 *Let f be in the form of (11). Then, under special regions of $|c_1|, |c_2|$ parameter space, f satisfies the DSMVC. Further, the union of these*

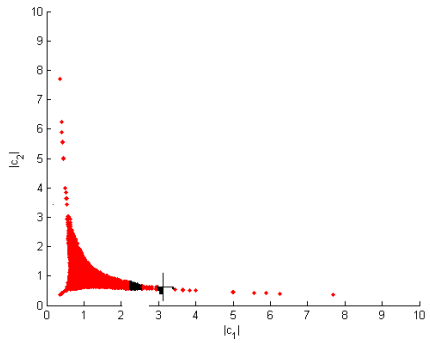


Figure 5: Red dots are values of $(|c_1|, |c_2|)$ where there exists c_1, c_2 such that the minimal critical point does not converge to the origin.

regions implies that the DSMVC is satisfied for all $|c_1|, |c_2|$ *except for possibly a compact set and where* f *is conjugate to a polynomial of form* $z - z^3 + a_4 z^4$.

To prove the theorem, we will first prove several regions of $|c_1|, |c_2|$ parameter space where f satisfies the DSMVC. We will then use these regions to show that the DSMVC is satisfied for all $|c_1|, |c_2|$, except for possibly a compact region in $|c_1|, |c_2|$ space: $[0, 8] \times [0, 8]$, and where f is conjugate to a polynomial of form $z - z^3 + a_4 z^4$.

We will first show the DSMVC holds in the following regions:

1. $|c_1| \geq 1.62$, $|c_2| \geq 1.62$
2. $|c_2| \leq 1.65$, $|c_1| \leq 0.23\,|c_2|$
3. $|c_2| \leq \frac{4}{5}$, $|c_1| \leq \frac{2}{5}\,|c_2|$
4. $\frac{2}{5}\,|c_1| \leq |c_2| \leq \frac{5}{2}\,|c_1|, |c_2| \leq \frac{4}{5} - |c_1|$

First, we conjugate f even further, by $\beta = \frac{1}{2}\left(\frac{1}{c_1} + \frac{1}{c_2} - c_1 c_2\right)$, to produce

$$h(z) := \beta f(\frac{z}{\beta}) = z - z^2 + Az^3 - Bz^4 \tag{12}$$

where

$$A = \frac{4c_1c_2\left(1 - c_1^2c_2 - c_1c_2^2\right)}{3\left(-c_1 - c_2 + c_1^2c_2^2\right)^2} \qquad B = \frac{2c_1^3c_2^3}{\left(-c_1 - c_2 + c_1^2c_2^2\right)^3}. \tag{13}$$

Note that unless $\beta = 0$, f is conjugate to a polynomial in the form $z - z^2 + a_3z^3 + a_4z^4$. If $\beta = 0$, then f is either conjugate to $z - z^3 + a_4z^4$ or f is conjugate to $z - z^4/4$. The second case is easy as we note 1 is a critical point of $z - z^4/4$, and it is easily checked 1 satisfies DSMVC. Thus, from here on forward, we will restrict our attention to quartics not conjugate to $z - z^3 + a_4z^4$.

7.1 $|c_1| \geq 1.62, |c_2| \geq 1.62$

When $|c_1| \geq 1.62$ and $|c_2| \geq 1.62$, $c_3 = -\dfrac{1}{c_1 c_2}$ satisfies the SMVC as the minimal critical point by Theorem 4.2. We will show that c_3 also converges to the origin under iteration of f.

From Lemmas 2.7 and 3.3, the conjugated map has critical point $v := \beta\left(-\dfrac{1}{c_1 c_2}\right)$, which converges to the origin under iteration of h exactly when c_3 converges to the origin under iteration of f.

Let v_4 denote the 4^{th} iterate of v under h. Using Mathematica, we composed h four times and expanded the expression into terms with powers of c_1 and c_2 using the ExpandAll command. This expression has thousands of terms. However, it starts with

$$v_4 = \frac{8463}{65536} - \frac{705815}{12582912 c_1^2 c_2} - \frac{705815}{12582912 c_1 c_2^2} - \frac{801767}{16777216 c_1^4 c_2^2} + \cdots$$

To bound the location of v_4, we use the triangle inequality to obtain:

$$\left| v_4 - \frac{8463}{65536} \right| \leq \left| \frac{705815}{12582912 c_1^2 c_2} \right| + \left| \frac{705815}{12582912 c_1 c_2^2} \right| + \left| \frac{801767}{16777216 c_1^4 c_2^2} \right| + \cdots \leq \frac{4}{100}.$$

To obtain this estimate, we used a text editor to change every minus to a plus, evaluated the expression at $c_1 = c_2 = \frac{162}{100}$, and bounded the expression from above.

Rounding $\frac{8463}{65536}$ up to 0.130, $v_4 \in D(\frac{8463}{65536}, 0.04) \subset D(0.85, 0.85) \subset D(0, 0.17)$. It follows that $|v_4| < 0.17$ and from Lemma 2.9, $Re\left(\dfrac{1}{v_4}\right) > \dfrac{100}{17}$. Next, we conjugate h by $\dfrac{1}{z} = w$ to produce

$$y(w) := \frac{1}{h\left(\dfrac{1}{z}\right)} = w + 1 + \frac{1 - A}{w} + \frac{B - 2A + 1}{w^2} + \frac{C}{w^3}, \tag{14}$$

where

$$C = \frac{(2B + A^2 - 3A + 1) + (B - A(2B - 2A + 1))z + (B(B - 2A + 1))z^2}{1 - z + Az^2 - Bz^3}. \tag{15}$$

Using the triangle inequality, we can bound $|A|$, $|B|$, and $|C|$ using $|c_1|, |c_2| \geq \frac{162}{100}$. Observe,

$$|A| = \left| \frac{4 c_1 c_2 \left(1 - c_1^2 c_2 - c_1 c_2^2\right)}{3\left(-c_1 - c_2 + c_1^2 c_2^2\right)^2} \right| \leq \frac{4\left(|c_1|\,|c_2| + |c_1|^3\,|c_2|^2 + |c_1|^2\,|c_2|^3\right)}{3\left(|c_1|^2\,|c_2|^2 - |c_1| - |c_2|\right)^2}. \tag{16}$$

Note that the bounds on c_1 are needed to use reverse triangle inequality on the denominator. We will now define $a : [1.62, \infty) \times [1.62, \infty) \to \mathbb{R}$ by

$$a(x, y) = \frac{\left(xy + x^3y^2 + x^2y^3\right)}{\left(x^2y^2 - x - y\right)^2}$$

and show that $a(x, y)$ has a maximum at $x = y = 1.62$ subject to the constraints $x, y \geq 1.62$.

Observe that

$$a_x(x, y) = \frac{xy - y^2 - x^3y^2 - 6x^2y^3 - 2xy^4 - x^4y^4 - 2x^3y^5}{\left(-x - y + x^2y^2\right)^3}$$

and

$$a_y(x, y) = \frac{-x^2 + xy - 2x^4y - 6x^3y^2 - x^2y^3 - 2x^5y^3 - x^4y^4}{\left(-x - y + x^2y^2\right)^3}.$$

Note that the first order partial derivatives exist on $[1.62, \infty) \times [1.62, \infty)$ and it is an exercise to show that

$$\nabla a(x, y) = 0 \Leftrightarrow (x, y) = ((-3)^{\frac{1}{3}}, (-3)^{\frac{1}{3}}) \notin [1.62, \infty) \times [1.62, \infty).$$

Thus the maximum of a occurs along the boundary. We will now consider the boundary in which $x = 1.62$. Note

$$a(1.62, y) = \frac{1.62y + 4.251528y^2 + 2.6244y^3}{\left(-1.62 - y + 2.6244y^2\right)^2}.$$

It is an exercise to show that $a(1.62, y)$ has no critical points with $y \in [1.62, \infty)$. Thus since $a_y(1.62, 1.62) < 0$, $a(1.62, y)$ is decreasing for $y \in [1.62, \infty)$ and thus the maximum occurs at $y = 1.62$. Since this argument is symmetric with respect to x and y, the maximum of a along the boundary occurs at $x = y = 1.62$. Thus it follows that

$$|A| \leq \frac{593941000000}{237627109443} < \frac{5}{2}.$$

Now bounding $|B|$ similarly, we obtain

$$|B| = \left|\frac{2c_1^3c_2^3}{\left(-c_1 - c_2 + c_1^2c_2^2\right)^3}\right| \leq \frac{2\left|c_1\right|^3\left|c_2\right|^3}{\left(\left|c_1\right|^2\left|c_2\right|^2 - \left|c_1\right| - \left|c_2\right|\right)^3}. \tag{17}$$

A similar argument shows that the maximum of this function of c_1 and c_2 occurs at $c_1 = c_2 = 1.62$. Thus

$$|B| \leq \frac{16607531250000000}{22292670436249121} < \frac{3}{4}.$$

Using the triangle inequality, we can bound $|C|$:

$$|C| = \left| \frac{(2B + A^2 - 3A + 1) + (B - A(2B - 2A + 1))z + (B(B - 2A + 1))}{1 - z + Az^2 - Bz^3} \right|$$

$$\leq \frac{(2|B| + |A|^2 + 3|A| + 1) + (|B| + 2|AB| + 2|A|^2 + |A|)|z| + (|B|^2 + 2|AB| + |B|)|z|^2}{1 - |z| - A|z|^2 - B|z|^3}. \tag{18}$$

Using the bounds for $|A|$ and $|B|$, we can further bound $|C|$ when $z \in D(0, 0.17)$:

$$|C| \leq \cdots \leq \frac{\frac{65}{4} + \frac{39}{2}|z| + \frac{81}{16}|z|^2}{1 - |z| - \frac{5}{2}|z|^2 - \frac{3}{4}|z|^3} \leq 26.2.$$

Let $w_4 = \dfrac{1}{v_4}$. Now, we will show that $Re(w_4)$ converges to $+\infty$ under iteration of y, which equivalently shows that v_4 converges to 0 under iteration of g.

We know that $|v_4| \leq 0.17 (\Leftrightarrow |w_4| \geq \frac{100}{17})$ and $Re(w_4) \geq \frac{100}{17}$, so the following inequalities hold under induction:

$$\begin{aligned} Re(y(w)) &\geq Re(w) + 1 - \frac{1 + |A|}{|w|} - \frac{|B| + 2|A| + 1}{|w|^2} - \frac{|C|}{|w|^3} \\ &\geq Re(w) + 1 - \frac{7}{2} \cdot 0.17 - \frac{27}{4} \cdot 0.17^2 - 26.2 \cdot 0.17^3 \\ &\geq Re(w) + 1 - 0.92 = Re(w) + 0.08. \end{aligned}$$

Thus the right half plane $Re(w) \geq \frac{100}{17}$ is forward-invariant under $y(w)$. Hence by induction, w_4 converges to ∞ under iteration of y, implying that v_4 converges to 0 under iteration of h.

The region where $|c_1|$ and $|c_2|$ satisfies $|c_1| \geq 1.62, |c_2| \geq 1.62$ is shown in Figure 6.

However, by showing the region $|c_1| \geq 1.62$, $|c_2| \geq 1.62$ satisfies the DSMVC, we have also shown several other symmetric regions satisfy the DSMVC. These regions are obtained from interchanging of the critical points involved in the inequalities:

1. $\left|\frac{1}{c_1 c_2}\right| \geq 1.62$, $|c_1| \geq 1.62$.
2. $\left|\frac{1}{c_1 c_2}\right| \geq 1.62$, $|c_2| \geq 1.62$.

All regions obtained from case $|c_1| \geq 1.62$, $|c_2| \geq 1.62$ are shown in Figure 7.

Figure 6: Region where $|c_1|, |c_2| \geq 1.62$.

Figure 7: All regions obtained from case $|c_1| \geq 1.62$, $|c_2| \geq 1.62$.

7.2 $|c_2| \leq 1.65$, $|c_1| \leq 0.23\,|c_2|$

From hypothesis, $|c_1|\,|c_2| \leq 0.23\,|c_2|^2 \leq 0.70$. Using $|c_1c_2c_3| = 1$, this implies $|c_3| > 1$. Thus, c_1 satisfies the SMVC as the minimal critical point by Theorem 4.2. We will show that c_1 also converges to the origin under iteration.

The conjugated map has critical point $u := \beta c_1 = \dfrac{1}{2} + \dfrac{c_1}{2c_2} - \dfrac{c_1^2 c_2}{2}$, which converges to the origin under iteration of h exactly when c_1 converges to the origin under iteration of f.

Let u_4 denote the 4^{th} iterate of u under h. Using Mathematica, we composed h four times and expanded the expression into terms with powers of c_1 and c_2 using the ExpandAll command. This expression has thousands of terms. However, it starts with

$$u_4 = \frac{8463}{65536} + \frac{469246883c_1^3}{6442450944} - \frac{31352311289c_1^6}{618475290624} + \frac{129811436695643c_1^9}{12824703626379264} + \cdots$$

Bounding the location of u_4 using the triangle inequality, we obtain

$$\left|u_4 - \frac{8463}{65536}\right| \leq \left|\frac{469246883c_1^3}{6442450944}\right| + \left|\frac{31352311289c_1^6}{618475290624}\right| + \left|\frac{129811436695643c_1^9}{12824703626379264}\right| + ...$$

Now since some terms have a power of c_2 in the denominator, we use the fact that $|c_1| \leq 0.23\,|c_2|$ to remove c_2 in the denominator of terms. We obtain

$$\left|u_4 - \frac{8463}{65536}\right| \leq \left|\frac{469246883(0.23c_2)^3}{6442450944}\right| + \left|\frac{31352311289(0.23c_2)^6}{618475290624}\right| + \cdots \leq \frac{4}{100}.$$

To obtain this estimate, we used a text editor to change every minus to a plus, evaluated the expression at $c_2 = 1.65$, and bounded the result from above.

So, $u_4 \in D(\frac{8463}{65536}, 0.04) \subset D(0.85, 0.85) \subset D(0, 0.17)$. It follows that $|u_4| \leq 0.17$ and from Lemma 2.9, $Re\left(\dfrac{1}{u_4}\right) \geq \dfrac{100}{17}$.

We will now consider $y(w)$ from (14) with A, B from (13) and C from (15). Using the triangle inequality, we can bound $|A|$, $|B|$, and $|C|$ using $|c_1| \leq 0.23 \cdot 1.65 = 0.3795$ and $|c_2| \leq 1.65$.

Bounding $|A|$ as in (16):

$$|A| \leq \frac{4\left(|c_1|\,|c_2| + |c_1|^3\,|c_2|^2 + |c_1|^2\,|c_2|^3\right)}{3\left(|c_2| - |c_1| - |c_1|^2\,|c_2|^2\right)^2}.$$

Note that the bounds on c_1 are needed to use reverse triangle inequality on the denominator. We will now define $a : T \to \mathbb{R}$ by

$$a(x, y) = \frac{\left(xy + x^3y^2 + x^2y^3\right)}{\left(y - x - x^2y^2\right)^2}$$

where $T = \{(x, y) \in \mathbb{R}^2 : 0 < y \leq 1.65 \text{ and } 0 < x \leq 0.23y\}$ and show that $a(x, y)$ has a maximum at $y = 1.65$, $x = 0.3795$. Observe that

$$a_x(x, y) = \frac{-xy - y^2 + x^3y^2 - 6x^2y^3 - 2xy^4 - x^4y^4 - 2x^3y^5}{(x - y + x^2y^2)^3}$$

and

$$a_y(x, y) = \frac{x^2 + xy + 2x^4y - x^2y^3 - 2x^5y^3 - x^4y^4}{(x - y + x^2y^2)^3}.$$

Note that the first order partial derivatives exist on T and it is an exercise to show that

$$\nabla a(x, y) = 0 \Leftrightarrow (x, y) = (3^{\frac{-2}{3}}, -3^{\frac{1}{3}}) \notin T.$$

Thus the maximum of a occurs along the boundary. Let L_1, L_2, and L_3 denote the lines connecting $(0, 0)$ to $(0, 1.65)$, $(0, 1.65)$ to $(0.3795, 1.65)$, and $(0.3795, 1.65)$ to $(0, 0)$ respectively.

1. L_1 has an equation of $x = 0$. Thus $a(0, y) = 0$ and the maximum of $a(x, y)$ along L_1 is 0.

2. L_2 has an equation of $y = 1.65$. Thus

$$a(x, 1.65) = \frac{1.65x + 4.492125x^2 + 2.7225x^3}{(1.65 - x - 2.7225x^2)^2}.$$

   It is an easy exercise to show that $a(x, 1.65)$ has no critical points with $x \in [0, 0.3795]$. Thus since $a_x(0, 1.65) > 0$, $a(x, 1.65)$ is increasing for $x \in [0, 0.3795]$ and thus the maximum occurs at $x = 0.3795$.

3. L_3 has an equation of $x = 0.23y$. Thus

$$a(0.23y, y) = \frac{0.23y^2 + 0.065067y^5}{(0.77y - 0.0529y^4)^2}.$$

   It is an easy exercise to show that $a(0.23y, y)$ has no critical point with $y \in (0, 1.65]$. Thus since $a_y(0.23(1.65), 1.65) > 0$, $a(0.23y, y)$ is increasing for $y \in (0, 1.65]$ and thus the maximum occurs at $y = 1.65$.

Notice that the maximum for L_2 and L_3 occurs at $(x, y) = (0.3795, 1.65)$ and it follows that

$$|A| \leq \frac{13370600892800000}{5441552322938787} \leq \frac{5}{2}.$$

Bounding $|B|$ as in (17), we obtain

$$|B| \leq \frac{2\,|c_1|^3\,|c_2|^3}{\left(|c_1|^2\,|c_2|^2 - |c_1| - |c_2|\right)^3}.$$

A similar argument shows that the maximum of this function of c_1 and c_2 occurs at $c_1 = 0.3795$ and $c_2 = 1.65$. Thus

$$|B| \leq \frac{42049152000000000000}{58039582085962835093} \leq \frac{3}{4}.$$

Bounding $|C|$ as in (18) and using the bounds for $|A|$ and $|B|$, we can bound $|C|$ when $z \in D(0, 0.17)$:

$$|C| \leq \cdots \leq \frac{\frac{65}{4} + \frac{39}{2}|z| + \frac{81}{16}|z|^2}{1 - |z| - \frac{5}{2}|z|^2 - \frac{3}{4}|z|^3} \leq 26.2.$$

Let $w_4 = \dfrac{1}{u_4}$. Now, we will show that $Re(w_4)$ converges to $+\infty$ under iteration of y, which equivalently shows that u_4 converges to 0 under iteration of g.
We know that $|u_4| \leq 0.17 (\Leftrightarrow |w_4| \geq \frac{100}{17})$ and $Re(w_4) \geq \frac{100}{17}$. Thus, the following inequalities hold under induction:

$$\begin{aligned} Re(y(w)) &\geq Re(w) + 1 - \frac{1 + |A|}{|w|} - \frac{|B| + 2|A| + 1}{|w|^2} - \frac{|C|}{|w|^3} \\ &\geq Re(w) + 1 - \frac{7}{2} \cdot 0.17 - \frac{27}{4} \cdot 0.17^2 - 26.2 \cdot 0.17^3 \\ &\geq Re(w) + 1 - 0.92 = Re(w) + 0.08. \end{aligned}$$

Thus the right half plane $Re(w) \geq \frac{100}{17}$ is forward-invariant under $y(w)$. Hence by induction, w_4 converges to ∞ under iteration of y, implying u_4 converges to 0 under iteration of h.

The region where $|c_2| \leq 1.65$, $|c_1| \leq 0.23|c_2|$ is shown in Figure 8.

However, by showing the region where $|c_2| \leq 1.65$, $|c_1| \leq 0.23|c_2|$ satisfies the DSMVC, we have actually shown the DSMVC is satisfied for symmetric regions resulting from interchanging of critical points involved in the inequalities:

1. $|c_2| \leq 1.65$, $\left|\frac{1}{c_1 c_2}\right| \leq 0.23|c_2|$
2. $|c_1| \leq 1.65$, $|c_2| \leq 0.23|c_1|$
3. $|c_1| \leq 1.65$, $\left|\frac{1}{c_1 c_2}\right| \leq 0.23|c_1|$
4. $\left|\frac{1}{c_1 c_2}\right| \leq 1.65$, $|c_1| \leq 0.23\left|\frac{1}{c_1 c_2}\right|$
5. $\left|\frac{1}{c_1 c_2}\right| \leq 1.65$, $|c_2| \leq 0.23\left|\frac{1}{c_1 c_2}\right|$

All the regions from case $|c_2| \leq 1.65$, $|c_1| \leq 0.23|c_2|$ are shown in Figure 9.

Figure 8: Region where $|c_2| \leq 1.65$, $|c_1| \leq 0.23\,|c_2|$.p

Figure 9: All the regions from case $|c_2| \leq 1.65$, $|c_1| \leq 0.23\,|c_2|$.

7.3 $|c_2| \leq \frac{4}{5}$, $|c_1| \leq \frac{2}{5}|c_2|$

Since $|c_2| \leq \frac{4}{5}$, we have $|c_1| \leq \frac{2}{5} \cdot \frac{4}{5} = \frac{8}{25}$, resulting in $|c_1||c_2| \leq \frac{32}{125}$. Using $|c_1c_2c_3| = 1$, $|c_3| > 1$ and thus c_1 satisfies the SMVC as the minimal critical point by Theorem 4.2. We will show that c_1 also converges to the origin under iteration.

The conjugated map has critical point $u := \beta c_1 = \frac{1}{2} + \frac{c_1}{2c_2} - \frac{c_1^2 c_2}{2}$, which converges to the origin under iteration of h exactly when c_1 converges to the origin under iteration of f.

Let u_4 denote the 4^{th} iterate of u under h. Using Mathematica, we composed h four times and expanded the expression into terms with powers of c_1 and c_2 using the ExpandAll command. This expression has thousands of terms. However, it starts with

$$u_4 = \frac{8463}{65536} + \frac{469246883c_1^3}{6442450944} - \frac{31352311289c_1^6}{618475290624} + \frac{129811436695643c_1^9}{12824703626379264} + \dots$$

Then, bounding the location of u_4 with the triangle inequality, we obtain

$$\left|u_4 - \frac{8463}{65536}\right| \leq \left|\frac{469246883c_1^3}{6442450944}\right| + \left|\frac{31352311289c_1^6}{618475290624}\right| + \left|\frac{129811436695643c_1^9}{12824703626379264}\right| + \cdots$$

Now since some terms have a power of c_2 in the denominator, we use the fact that $|c_1| \leq \frac{2}{5}|c_2|$ to remove c_2 in the denominator. We obtain

$$\left|u_4 - \frac{8463}{65536}\right| \leq \left|\frac{469246883(\frac{2}{5}c_2)^3}{6442450944}\right| + \left|\frac{31352311289(\frac{2}{5}c_2)^6}{618475290624}\right| + \cdots \leq \frac{401}{10000}.$$

To obtain this estimate, we used a text editor to change every minus to a plus, evaluated the expression at $c_2 = \frac{4}{5}$, and bounding the result from above. Rounding $\frac{8463}{65536}$ up to 0.130, $u_4 \in D(\frac{8463}{65536}, 0.0401) \subset D(0.085, 0.085) \subset D(0, 0.17)$. It follows that $|u_4| \leq 0.17$ and from Lemma 2.9, $Re\left(\frac{1}{u_4}\right) \geq \frac{100}{17}$.

We will now consider $y(w)$ from (14) as defined above with A, B from (13) and C from (15). Using the triangle inequality, we can bound $|A|$, $|B|$, and $|C|$ using $|c_1| \leq \frac{8}{25}$ and $|c_2| \leq \frac{4}{5}$.

Bounding $|A|$ as in (16):

$$|A| \leq \frac{4\left(|c_1||c_2| + |c_1|^3|c_2|^2 + |c_1|^2|c_2|^3\right)}{3\left(|c_2| - |c_1| - |c_1|^2|c_2|^2\right)^2}.$$

Note that the bounds on c_1 are needed to use reverse triangle inequality on the denominator. We will now define $a : T \to \mathbb{R}$ by

$$a(x, y) = \frac{\left(xy + x^3y^2 + x^2y^3\right)}{\left(y - x - x^2y^2\right)^2}$$

where $T = \{(x, y) \in \mathbb{R}^2 : 0 < y \leq \frac{4}{5} \text{ and } 0 < x \leq \frac{2}{5}y\}$ and show that $a(x, y)$ has a maximum at $y = \frac{4}{5}$, $x = \frac{8}{25}$. Observe that

$$a_x(x, y) = \frac{-xy - y^2 + x^3y^2 - 6x^2y^3 - 2xy^4 - x^4y^4 - 2x^3y^5}{(x - y + x^2y^2)^3}$$

and

$$a_y(x, y) = \frac{x^2 + xy + 2x^4y - x^2y^3 - 2x^5y^3 - x^4y^4}{(x - y + x^2y^2)^3}.$$

Note that the first order partial derivatives exist on T and it is an exercise to show that

$$\nabla a(x, y) = 0 \Leftrightarrow (x, y) = (3^{\frac{-2}{3}}, -3^{\frac{1}{3}}) \notin T.$$

Thus the maximum of a occurs along the boundary. Let L_1, L_2, and L_3 denote the lines connecting $(0, 0)$ to $(0, \frac{4}{5})$, $(0, \frac{4}{5})$ to $(\frac{8}{25}, \frac{4}{5})$, and $(\frac{8}{25}, \frac{4}{5})$ to $(0, 0)$ respectively.

1. L_1 has an equation of $x = 0$. Thus $a(0, y) = 0$ and the maximum of $a(x, y)$ along L_1 is 0.

2. L_2 has an equation of $y = \frac{4}{5}$. Thus

$$a\left(x, \frac{4}{5}\right) = \frac{\frac{4x}{5} + \frac{64x^2}{125} + \frac{16x^3}{25}}{\left(\frac{4}{5} - x - \frac{16x^2}{25}\right)^2}.$$

   It is an easy exercise to show that $a(x, \frac{4}{5})$ has no critical points with $x \in [0, \frac{8}{25}]$. Thus since $a_x(0, \frac{4}{5}) > 0$, $a(x, \frac{4}{5})$ is increasing for $x \in [0, \frac{8}{25}]$ and thus the maximum occurs at $x = \frac{8}{25}$.

3. L_3 has an equation of $x = \frac{2}{5}y$. Thus

$$a\left(\frac{2}{5}y, y\right) = \frac{\frac{4y^2}{5} + \frac{144y^5}{125}}{\left(\frac{y}{5} - \frac{16y^4}{25}\right)^2}.$$

   It is an easy exercise to show that $a(\frac{2}{5}y, y)$ has no critical point with $y \in (0, \frac{4}{5}]$. Thus since $a_y(\frac{2}{5}(\frac{4}{5}), \frac{4}{5}) > 0$, $a(\frac{2}{5}, y)$ is increasing for $y \in (0, \frac{4}{5}]$ and thus the maximum occurs at $y = \frac{4}{5}$.

Notice that the maximum for L_2 and L_3 occurs at $(x, y) = (\frac{8}{25}, \frac{4}{5})$ and it follows that

$$|A| \leq \frac{20105000}{7863483} \leq \frac{13}{5}.$$

Bounding $|B|$ as in (17),

$$|B| \leq \frac{2\left|c_1\right|^3\left|c_2\right|^3}{\left(\left|c_2\right|-\left|c_1\right|-\left|c_1\right|^2\left|c_2\right|^2\right)^3}.$$

A similar argument shows that the maximum of this function of c_1 and c_2 occurs at $c_1 = \frac{8}{25}$ and $c_2 = \frac{4}{5}$. Thus

$$|B| \leq \frac{2000000000}{4243659659} \leq \frac{1}{2}.$$

Bounding $|C|$ as in (18) and using the bounds for $|A|$ and $|B|$, we can bound $|C|$ when $z \in D(0, 0.17)$:

$$|C| \leq \cdots \leq \frac{\frac{414}{25} + \frac{961}{50}|z| + \frac{67}{20}|z|^2}{1 - |z| - \frac{13}{5}|z|^2 - \frac{1}{2}|z|^3} \leq 26.5.$$

Let $w_4 = \dfrac{1}{u_4}$. Now, we will show that $Re(w_4)$ converges to $+\infty$ under iteration of y, which equivalently shows that u_4 converges to 0 under iteration of g.
We know that $|u_4| \leq 0.17 (\Leftrightarrow |w_4| \geq \frac{100}{17})$ and $Re(w_4) \geq \frac{100}{17}$. Thus, the following inequalities hold under induction:

$$\begin{aligned} Re(y(w)) &\geq Re(w) + 1 - \frac{1+|A|}{|w|} - \frac{|B|+2|A|+1}{|w|^2} - \frac{|C|}{|w|^3} \\ &\geq Re(w) + 1 - \frac{18}{5}\cdot 0.17 - \frac{67}{10}\cdot 0.17^2 - 26.5\cdot 0.17^3 \\ &\geq Re(w) + 1 - 0.94 = Re(w) + 0.06. \end{aligned}$$

Thus the right half plane $Re(w) \geq \frac{100}{17}$ is forward-invariant under $y(w)$. Hence by induction, w_4 converges to ∞ under iteration of y, implying that u_4 converges to 0 under iteration of h.

The region where $|c_2| \leq \frac{4}{5}$, $|c_1| \leq \frac{2}{5}|c_2|$ is shown in Figure 10.

Note that we have actually proved more than just the region $|c_2| \leq \frac{4}{5}$, $|c_1| \leq \frac{2}{5}|c_2|$ satisfies the DSMVC. There are symmetric regions resulting from interchanging of critical points in the inequality that also satisfy the DSMVC:

1. $|c_2| \leq \frac{4}{5}$, $\left|\frac{1}{c_1 c_2}\right| \leq \frac{2}{5}|c_2|$
2. $|c_1| \leq \frac{4}{5}$, $|c_2| \leq \frac{2}{5}|c_1|$
3. $|c_1| \leq \frac{4}{5}$, $\left|\frac{1}{c_1 c_2}\right| \leq \frac{2}{5}|c_1|$
4. $\left|\frac{1}{c_1 c_2}\right| \leq \frac{4}{5}$, $|c_1| \leq \frac{2}{5}\left|\frac{1}{c_1 c_2}\right|$
5. $\left|\frac{1}{c_1 c_2}\right| \leq \frac{4}{5}$, $|c_2| \leq \frac{2}{5}\left|\frac{1}{c_1 c_2}\right|$

Figure 10: Region where $|c_2| \leq \frac{4}{5}$, $|c_1| \leq \frac{2}{5}|c_2|$.p

All regions obtained from the case $|c_2| \leq \frac{4}{5}$, $|c_1| \leq \frac{2}{5}|c_2|$ are shown in Figure 11.

7.4 $\frac{2}{5}|c_1| \leq |c_2| \leq \frac{5}{2}|c_1|, |c_2| \leq \frac{4}{5} - |c_1|$

In this section, we will prove a region of the SMVC using a different technique than before. We will show for all monic polynomials f, there exists an R such that $|z| > R \Rightarrow |f(z)| > R$. Using this, we will provide a sufficient condition of one of the critical points always "staying away" from the origin under iteration and the other two critical points satisfying the SMVC. Note that this implies the DSMVC, as one critical point that satisfies the SMVC always converges to the origin.

We will first prove the following:

Lemma 7.2 *Let* $p(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_{n-1} z^{n-1} + z^n$, *where* $a_0, \ldots, a_{n-1} \in \mathbb{C}$. *Let* $R = 1 + \sum_{i=0}^{n-1} |a_i|$. *If* $|z| > R$, *then* $|p(z)| > R$, $\forall z \in \mathbb{C}$.

Figure 11: All regions obtained from case $|c_2| \leq \frac{4}{5}$, $|c_1| \leq \frac{2}{5}|c_2|$

Proof For $|z| > R > 1$ the following inequalities prove the result:

$$\begin{aligned}
|p(z)| &\geq |z|^n - \left|a_{n-1}z^{n-1} + \cdots + a_2z^2 + a_1z + a_0\right| \\
&\geq |z|^n - \sum_{i=0}^{n-1} |a_i|\,|z|^i \geq |z|^n - |z|^{n-1}\sum_{i=0}^{n-1}|a_i| = |z|^{n-1}\left(|z| - \sum_{i=0}^{n-1}|a_i|\right) \\
&\geq |z|\left(|z| - \sum_{i=0}^{n-1}|a_i|\right) > \left(1 + \sum_{i=0}^{n-1}|a_i|\right)\left(1 + \sum_{i=0}^{n-1}|a_i| - \sum_{i=0}^{n-1}|a_i|\right) \\
&= 1 + \sum_{i=0}^{n-1}|a_i| = R.
\end{aligned}$$

□

Now, we will further conjugate f from (11) by $\alpha = \frac{1}{\sqrt[3]{4}}$ so we can consider the monic polynomial

$$\tilde{f}(z) = \alpha f(\frac{z}{\alpha}) = z + \frac{\left(-c_1 - c_2 + c_1^2c_2^2\right)}{2^{1/3}c_1c_2}z^2 - \frac{2^{4/3}\left(-1 + c_1^2c_2 + c_1c_2^2\right)}{3c_1c_2}z^3 + z^4.$$

The critical points of $\tilde{f}$ are $\frac{c_1}{\sqrt[3]{4}}$, $\frac{c_1}{\sqrt[3]{4}}$, and $-\frac{1}{\sqrt[3]{4}c_1c_2}$.

Now, we will prove the following result:

Lemma 7.3 *A region where* $-\dfrac{1}{c_1c_2}$ *does not converge to the origin under iter-*

ation of f is defined by the inequality:

$$\frac{1-2\left|c_1\right|^2\left|c_2\right|-2\left|c_1\right|\left|c_2\right|^2-6\left|c_1\right|^3\left|c_2\right|^3}{12\cdot 2^{2/3}\left|c_1\right|^4\left|c_2\right|^4}$$
$$>2+\frac{\left|c_1\right|+\left|c_2\right|+\left|c_1\right|^2\left|c_2\right|^2}{2^{1/3}\left|c_1\right|\left|c_2\right|}+\frac{2^{4/3}\left(1+\left|c_1\right|^2\left|c_2\right|+\left|c_1\right|\left|c_2\right|^2\right)}{3\left|c_1\right|\left|c_2\right|}. \quad (19)$$

Proof

$$R=1+\sum_{i=0}^{n-1}|a_i|=1+1+\left|\frac{\left(-c_1-c_2+c_1^2c_2^2\right)}{2^{1/3}c_1c_2}\right|+\left|\frac{2^{4/3}\left(-1+c_1^2c_2+c_1c_2^2\right)}{3c_1c_2}\right|$$
$$\leq 2+\frac{\left|c_1\right|+\left|c_2\right|+\left|c_1\right|^2\left|c_2\right|^2}{2^{1/3}\left|c_1\right|\left|c_2\right|}+\frac{2^{4/3}\left(1+\left|c_1\right|^2\left|c_2\right|+\left|c_1\right|\left|c_2\right|^2\right)}{3\left|c_1\right|\left|c_2\right|}.$$

Note that by Lemma 2.7, $-\alpha\dfrac{1}{c_1c_2}=-\dfrac{1}{\sqrt[3]{4}c_1c_2}$ is a critical point of $\tilde{f}$. The first iterate of $-\dfrac{1}{\sqrt[3]{4}c_1c_2}$ under $\tilde{f}$ is:

$$v_1:=-\frac{1+2c_1^2c_2+2c_1c_2^2+6c_1^3c_2^3}{12\cdot 2^{2/3}c_1^4c_2^4}.$$

Observe,

$$|v_1|=\left|\frac{1+2c_1^2c_2+2c_1c_2^2+6c_1^3c_2^3}{12\cdot 2^{2/3}c_1^4c_2^4}\right|\geq\frac{1-2\left|c_1\right|^2\left|c_2\right|-2\left|c_1\right|\left|c_2\right|^2-6\left|c_1\right|^3\left|c_2\right|^3}{12\cdot 2^{2/3}\left|c_1\right|^4\left|c_2\right|^4}.$$

Thus by Lemma 7.2,

$$\frac{1-2\left|c_1\right|^2\left|c_2\right|-2\left|c_1\right|\left|c_2\right|^2-6\left|c_1\right|^3\left|c_2\right|^3}{12\cdot 2^{2/3}\left|c_1\right|^4\left|c_2\right|^4}$$
$$>2+\frac{\left|c_1\right|+\left|c_2\right|+\left|c_1\right|^2\left|c_2\right|^2}{2^{1/3}\left|c_1\right|\left|c_2\right|}+\frac{2^{4/3}\left(1+\left|c_1\right|^2\left|c_2\right|+\left|c_1\right|\left|c_2\right|^2\right)}{3\left|c_1\right|\left|c_2\right|}.$$

implies $|v_1|>R$, which in turn through induction implies

$$\left|\lim_{n\to\infty}\tilde{f}^n(-\frac{1}{\sqrt[3]{4}c_1c_2})\right|>R.$$

Thus by Lemma 3.3, $-\dfrac{1}{c_1c_2}$ does not converge to the origin under iteration of f.

□

Now, we will show we can find a region inside the region defined by (19) that has a much simpler implicit function with respect to $|c_1|$ and $|c_2|$.

Lemma 7.4 *The bound* $|c_2| \leq \frac{4}{5} - |c_1|$ *implies the inequality (19) from Lemma 7.3 and hence* $-\dfrac{1}{c_1 c_2}$ *does not converge to the origin under iteration of* f *when* $|c_2| \leq \frac{4}{5} - |c_1|$.

Proof The inequality (19) from Lemma 7.3 is equivalent to

$$\begin{aligned} 2^{4/3}|c_1|^2|c_2| + 2^{4/3}|c_1||c_2|^2 + 22\cdot 2^{1/3}|c_1|^3|c_2|^3 \\ + 12\cdot 2^{2/3}|c_1|^4|c_2|^3 + 12\cdot 2^{2/3}|c_1|^3|c_2|^4 + 48|c_1|^4|c_2|^4 \\ + 16\cdot 2^{1/3}|c_1|^5|c_2|^4 + 16\cdot 2^{1/3}|c_1|^4|c_2|^5 + 12\cdot 2^{2/3}|c_1|^5|c_2|^5 \\ < 2^{1/3}. \end{aligned}$$

Let ♣ be the left-hand side of the previous inequality.

As the trivial bounds of $|c_1|, |c_2| \leq 0.8$ are not strong enough, we will bound $|c_1||c_2|$. Using the inequality of arithmetic and geometric means, we have for any $|c_1|, |c_2|$:

$$|c_1||c_2| \leq \left(\frac{|c_1|+|c_2|}{2}\right)^2 \leq \left(\frac{2}{5}\right)^2 = \frac{4}{25}.$$

Now, applying our new bound of $|c_1||c_2|$, we obtain the following inequalities that prove the result:

$$\begin{aligned} \clubsuit &\leq 2^{4/3}\cdot\frac{4}{5}\cdot\frac{4}{25} + 2^{4/3}\cdot\frac{4}{5}\cdot\frac{4}{25} + 22\cdot 2^{1/3}\cdot\left(\frac{4}{25}\right)^3 \\ &+ 12\cdot 2^{2/3}\left(\frac{4}{25}\right)^3\frac{4}{5} + 12\cdot 2^{2/3}\left(\frac{4}{25}\right)^3\frac{4}{5} + 48\left(\frac{4}{25}\right)^4 \\ &+ 16\cdot 2^{1/3}\left(\frac{4}{25}\right)^4\frac{4}{5} + 16\cdot 2^{1/3}\left(\frac{4}{25}\right)^4\frac{4}{5} + 12\cdot 2^{2/3}\left(\frac{4}{25}\right)^5 \\ &= \frac{12288}{390625} + \frac{1208768\cdot 2^{1/3}}{1953125} + \frac{780288\cdot 2^{2/3}}{9765625} < 1 < 2^{1/3}. \end{aligned}$$

Thus it follows that if $|c_2| \leq \frac{4}{5} - |c_1|$, then $-\dfrac{1}{c_1 c_2}$ does not converge to the origin under iteration of f. □

Lemma 7.5 *If* $\frac{2}{5}|c_1| \leq |c_2| \leq \frac{5}{2}|c_1|$ *and* $|c_2| \leq \frac{4}{5} - |c_1|$, *then* c_1 *and* c_2 *satisfy the SMVC.*

Proof We will first find a different set of sufficient conditions for c_1 and c_2 to satisfy the SMVC, then show the conditions laid out in the lemma imply those conditions.

It is clear that

$$\left|\frac{f(c_1)}{c_1}\right| \leq 1 \Leftrightarrow \left|-6 - c_1^3 + \frac{2c_1}{c_2} + 2c_1^2 c_2\right| \leq 12.$$

Using the triangle inequality, we see that a sufficient condition is:

$$|c_1|^3 + \frac{2|c_1|}{|c_2|} + 2|c_1|^2|c_2| \leq 6. \tag{20}$$

It is also clear that

$$\left|\frac{f(c_2)}{c_2}\right| \leq 1 \Leftrightarrow \left|-6 - c_2^3 + \frac{2c_2}{c_1} + 2c_2^2 c_1\right| \leq 12.$$

Using the triangle inequality, we see that a sufficient condition is:

$$|c_2|^3 + \frac{2|c_2|}{|c_1|} + 2|c_2|^2|c_1| \leq 6. \tag{21}$$

Now, we will prove that $\frac{2}{5}|c_1| \leq |c_2| \leq \frac{5}{2}|c_1|$ and $|c_2| \leq \frac{4}{5} - |c_1|$ imply (20). This will also imply (21) as the entire argument is symmetric about c_1 and c_2. It is easy to see that the maximum of $|c_1|$ and $|c_2|$ in the region defined by the inequalities in the hypothesis is the value of $|c_2|$ at the intersection of the lines $|c_2| = \frac{4}{5} - |c_1|$ and $|c_2| = \frac{5}{2}|c_1|$, which is $\frac{4}{7}$.

It is easy to see that (20) is equivalent to

$$|c_1|^3|c_2| + 2|c_1| + 2|c_1|^2|c_2|^2 \leq 6|c_2|.$$

The following chain of inequalities proves the result:

$$|c_1|^3|c_2| + 2|c_1| + 2|c_1|^2|c_2|^2 \leq \left(\frac{4}{7}\right)^3 |c_2| + 2{\cdot}2.5|c_2| + 2\left(\frac{4}{7}\right)^3 |c_2| = \frac{1907}{343}|c_2| < 6|c_2|.$$

□

Using Lemmas 7.5 and 7.4, we see that the region $N = \{(|c_1|, |c_2|) : \frac{2}{5}|c_1| \leq |c_2| \leq \frac{5}{2}|c_1|, |c_2| \leq \frac{4}{5} - |c_1|\}$ satisfies the DSMVC. The region N is shown in Figure 12. However, there are also symmetric regions that are also taken care of from relabeling of critical points:

1. $|c_2| \geq \frac{2}{5|c_1||c_2|}$, $\frac{1}{|c_1||c_2|} \geq \frac{2|c_2|}{5}, |c_2| \leq \frac{4}{5} - \frac{1}{|c_1||c_2|}$
2. $|c_1| \geq \frac{2}{5|c_1||c_2|}$, $\frac{1}{|c_1||c_2|} \geq \frac{2|c_1|}{5}, |c_1| \leq \frac{4}{5} - \frac{1}{|c_1||c_2|}$

All of these regions are shown in Figure 13.

7.5 Compact region

We will conclude this section by showing that the DSMVC is satisfied except possibly for $[0, 8] \times [0, 8]$ in $|c_1|, |c_2|$ space.

We will first simplify the inequalities defining a region that was proved in the previous section to satisfy the SMVC:

Figure 12: Region defined by $\frac{2}{5}|c_1| \leq |c_2| \leq \frac{5}{2}|c_1|$ and $|c_2| \leq \frac{4}{5} - |c_1|$.

Figure 13: All regions from case $\frac{2}{5}|c_1| \leq |c_2| \leq \frac{5}{2}|c_1|, |c_2| \leq \frac{4}{5} - |c_1|$.

Lemma 7.6 *The region*

$$S = \{(|c_1|, |c_2|) : |c_2| \geq \frac{2}{5|c_1||c_2|},\ \frac{1}{|c_1||c_2|} \geq \frac{2|c_2|}{5}, |c_2| \leq \frac{4}{5} - \frac{1}{|c_1||c_2|}\}$$

is defined by

$$\sqrt{\frac{2}{5|c_1|}} \leq |c_2| \leq \sqrt{\frac{5}{2|c_1|}}$$

for $|c_1| \geq 8$.

Proof Algebra shows the following equivalences:

$$|c_2| \geq \frac{2}{5|c_1||c_2|} \Leftrightarrow |c_2| \geq \sqrt{\frac{2}{5|c_1|}}$$

$$\frac{1}{|c_1||c_2|} \geq \frac{2|c_2|}{5} \Leftrightarrow |c_2| \leq \sqrt{\frac{5}{2|c_1|}}$$

$$|c_2| \leq \frac{4}{5} - \frac{1}{|c_1||c_2|} \Leftrightarrow \frac{2}{5} - \frac{1}{5}\sqrt{\frac{-25+4|c_1|}{|c_1|}} \leq |c_2| \leq \frac{2}{5} + \frac{1}{5}\sqrt{\frac{-25+4|c_1|}{|c_1|}}.$$

To prove the desired result, it is enough to show that the following inequalities hold when $|c_1| \geq 8$:

$$\sqrt{\frac{5}{2|c_1|}} \leq \frac{2}{5} + \frac{1}{5}\sqrt{\frac{-25+4|c_1|}{|c_1|}} \tag{22}$$

$$\sqrt{\frac{2}{5|c_1|}} \geq \frac{2}{5} - \frac{1}{5}\sqrt{\frac{-25+4|c_1|}{|c_1|}}. \tag{23}$$

Define the following functions:
$\phi : [\frac{25}{4}, \infty) \to \mathbb{R}$, where

$$\phi(x) = \frac{2}{5} + \frac{1}{5}\sqrt{\frac{-25+4x}{x}} - \sqrt{\frac{5}{2x}} = \frac{-5\sqrt{10} + 4\sqrt{x} + 2\sqrt{-25+4x}}{10\sqrt{x}}$$

$\psi : [\frac{25}{4}, \infty) \to \mathbb{R}$, where

$$\psi(x) = \sqrt{\frac{2}{5x}} - \left(\frac{2}{5} - \frac{1}{5}\sqrt{\frac{-25+4x}{x}}\right) = \frac{\sqrt{10} - 2\sqrt{x} + \sqrt{-25+4x}}{5\sqrt{x}}.$$

We will use ϕ and ψ to show (22) and (23) hold for $|c_1| \geq 8$.
Note that

$$\phi(|c_1|) \geq 0 \Leftrightarrow (22) \text{ holds.}$$

$$\psi(|c_1|) \geq 0 \Leftrightarrow (23) \text{ holds.}$$

Algebra shows that $\frac{245}{32}$ is the only zero of ϕ. It is easy to see $\frac{245}{32} < 8$. Since ϕ is continuous on $[8, \infty)$ and $\phi(8) = \frac{2}{5} + \frac{\sqrt{\frac{7}{2}}}{10} - \frac{\sqrt{5}}{4} > 0$, it follows that $\phi(x) > 0$ for $x \geq 8$. Hence (22) holds for $|c_1| \geq 8$. Algebra shows that $\frac{245}{32}$ is the only zero of ψ. It is easy to see $\frac{245}{32} < 8$. Since ψ is continuous on $[8, \infty)$ and $\psi(8) = -\frac{2}{5} + \frac{\sqrt{\frac{7}{2}}}{10} + \frac{1}{2\sqrt{5}} > 0$, it follows that $\psi(x) > 0$ for $x \geq 8$. Hence (23) holds for $|c_1| \geq 8$. □

Now, we will prove the main theorem of the section: showing that the DSMVC is satisfied except for possibly on a compact set in $|c_1|$, $|c_2|$ parameter space.

Theorem 7.7 *Let f be as in (11). Then, if c_1 and c_2 satisfy at least one of the following:*

1. $|c_1| \geq 8$
2. $|c_2| \geq 8$
3. $\frac{1}{|c_1|\,|c_2|} \geq 8$,

then the DSMVC holds.

Proof We will prove that the DSMVC holds for $|c_1| \geq 8$. The other two cases follow trivially from interchanging the labeling of critical points.

Suppose that $|c_1| \geq 8$. From previous results, we know the conjecture holds for

1. $0 < |c_2| \leq \frac{50}{81\,|c_1|}$ (from case $|c_1| \geq 1.62, |c_2| \geq 1.62$)
2. $\frac{20}{33\,|c_1|} \leq |c_2| \leq \sqrt{\frac{23}{100\,|c_1|}}$ (from case $|c_2| \leq 1.65$, $|c_1| \leq 0.23\,|c_2|$)
3. $\frac{5}{4\,|c_1|} \leq |c_2| \leq \sqrt{\frac{2}{5\,|c_1|}}$ (from case $|c_2| \leq \frac{4}{5}$, $|c_1| \leq \frac{2}{5}\,|c_2|$)
4. $\sqrt{\frac{2}{5\,|c_1|}} \leq |c_2| \leq \sqrt{\frac{5}{2\,|c_1|}}$ (from Lemma 7.6)
5. $\sqrt{\frac{5}{2\,|c_1|}} \leq |c_2| \leq \frac{4}{5}$ (from case $|c_2| \leq \frac{4}{5}$, $|c_1| \leq \frac{2}{5}\,|c_2|$)
6. $\sqrt{\frac{100}{23\,|c_1|}} \leq |c_2| \leq \frac{33}{20}$ (from case $|c_2| \leq 1.65$, $|c_1| \leq 0.23\,|c_2|$)

7. $\frac{81}{50} \leq |c_2|$ (from case $|c_1| \geq 1.62, |c_2| \geq 1.62$)

Thus, to show for $|c_1| \geq 8$, all $(|c_1|, |c_2|)$ is in a region that satisfies the DSMVC, it suffices to show the following inequalities hold for $|c_1| \geq 8$:

$$\frac{20}{33\,|c_1|} \leq \frac{50}{81\,|c_1|} \tag{24}$$

$$\frac{5}{4\,|c_1|} \leq \sqrt{\frac{23}{100\,|c_1|}} \tag{25}$$

$$\sqrt{\frac{100}{23\,|c_1|}} \leq \frac{4}{5} \tag{26}$$

$$\frac{81}{50} \leq \frac{33}{20} \tag{27}$$

Note that (24) and (27) are trivially true.
Define the following functions:
$\phi : \mathbb{R}^+ \to \mathbb{R}$, where

$$\phi(x) = \sqrt{\frac{23}{100x}} - \frac{5}{4x} = \frac{2\sqrt{23x} - 25}{20x}.$$

$\psi : \mathbb{R}^+ \to \mathbb{R}$, where

$$\psi(x) = \frac{4}{5} - \sqrt{\frac{100}{23x}}.$$

It is easy to see that

$$\phi(|c_1|) \geq 0 \Leftrightarrow (25) \text{ holds}$$
$$\psi(|c_1|) \geq 0 \Leftrightarrow (26) \text{ holds.}$$

Algebra shows that the only zero of ϕ is $\frac{25^2}{4(23)} \notin [8, \infty)$. Since ϕ is continuous on $[8, \infty)$ and $\phi(8) > 0$, it is clear that $\phi(x) \geq 0$, $\forall x \in [8, \infty)$. Thus, (25) holds for $|c_1| \geq 8$. Calculation shows that $\psi'(x) = \frac{1}{2}\sqrt{\frac{100}{23}}\frac{1}{x^{\frac{3}{2}}}$. Thus, ψ is increasing on $[8, \infty)$. Since $\psi(8) = \frac{4}{5} - \sqrt{\frac{100}{23 \cdot 8}} > 0$, then $\psi(x) \geq 0$, $\forall x \in [8, \infty)$. Thus, (26) holds for $|c_1| \geq 8$.

□

8 Acknowledgements

We would like to thank Professor Kevin Pilgrim for mentoring us over the past 8 weeks, and for organizing the REU as a whole.

Classification of Critically Fixed Rational Functions

Nicholas Nuechterlein *Samantha Pinella**

Abstract

In 1989, Tischler showed that the set of all complex polynomials whose critical points are also fixed points can be enumerated explicitly as branched mappings [1]. Much research has been done to do the same for rational functions $f : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ where $\hat{\mathbb{C}} := \mathbb{C} \cup \{\infty\}$, $f(z) = p(z)/q(z)$, $\gcd(p, q) = 1$ and the degree of f is $d := \max\{\deg(p), \deg(q)\}$. Nekrashevych showed that the dynamics of such maps f are determined by an algebraic invariant called a *wreath recursion* on a free group G, a homomorphism $\Phi : G \to G^d \rtimes S_d$ [2]. Motivated by a newly developed computer program by Bartholdi which finds numerical approximations of f given a wreath recursion, we develop an algorithm to compute the wreath recursions of branched mappings on $\hat{\mathbb{C}}$.

1 Introduction

A natural question to ask in complex dynamical systems is how we might classify self mappings of the extended complex plane, $\hat{\mathbb{C}} := \mathbb{C} \cup \{\infty\}$. In what follows we consider this question in the context of rational functions $f(z) = p(z)/q(z)$ whose critical points are all fixed. Here p, q are polynomials and $\gcd(p, q) = 1$. By *critical* we mean not locally injective: such functions f have $2d - 2$ critical points, where $d := \max\{\deg(p), \deg(q)\}$, counting with multiplicity. In 1989, Tischler used Thurston's topological characterization of critically finite rational mappings to classify all complex polynomials whose critical points are also fixed [1]. Naturally, we ask if a similar method using branched mappings can be used to classify critically fixed rational functions. Nekrashevych showed that, up to conjugation by Möbius transformations, the dynamics of such maps f are determined by an algebraic invariant called a *wreath recursion* on a free group G, where G is the fundamental group of $\hat{\mathbb{C}} \setminus \mathrm{Crit}(f)$ and a wreath recursion is a homomorphism $\Phi : G \to G^d \rtimes S_d$ [2]. Here S_d acts on G^d by permuting its coordinates.

A wreath recursion of a rational function is most conveniently found using a topological description of it as a branched mapping. A 'blowing up' construction of multigraphs on the 2-sphere will provide some branched mappings. A 2008

*The authors would like to thank the NSF for their support of the Indiana University REU.

result of Liu and Osserman implies that any rational map with fixed critical points is a *twist* of a blown up graph [3]. In [4], it is shown that any admissible partition of critical points can be represented as the degree set of a multigraph.

Our work is motivated by a recently developed computer program by Bartholdi which provides numerical approximations of rational functions, given their wreath recursions. In this paper, we describe a process of finding such wreath recursions and conclude by stating and proving an algorithm for doing so. We start by reviewing some necessary background information.

2 Background

2.1 Definitions

We will need a few basic concepts from algebraic topology, most importantly (1) the fundamental group $\pi_1(X, b)$ of a topological space X based at a point $b \in X$, (2) the notion of a covering $p : \tilde{X} \to X$ of X, and (3) the monodromy action of the fundamental group $\pi_1(X, b)$ on the preimage $p^{-1}(b)$ of the basepoint b under the covering map p.

Definition: The fundamental group $\pi_1(X, b)$ of a topological space X based at $b \in X$ is the set of all homotopy classes of loops in X based at b. Composition is the concatenation of paths. The identity is the constant path.

Informally speaking, two paths are homotopic if they can be continuously deformed into each other: intuitively we may think of this as stretching paths. If $[f]$ and $[g]$ are two homotopy classes of paths, we define $[f] \circ [g] = [f * g]$ where $f * g$ means one traverses the path f first then g. For obvious reasons, if a path starts and ends at the same point we call it a loop.

Definition: We say a space $\tilde{X}$ together with a map $p : \tilde{X} \to X$ is a *covering space* if for each $x \in X$ there exists a neighborhood U_x such that

$$p^{-1}(U_x) = \coprod_{j \in J} V_{x,j} \qquad \text{and} \qquad p|_{V_{x,j}} : V_{x,j} \longrightarrow U_x$$

is a homeomorphism for each $j \in J$.

Before we define the monodromy action, we introduce the notion of a path lift. Suppose $\omega : [0, 1] \to X$ is a loop in X based at b. The *path lift* $\tilde{\omega}_{b_j}$ of ω based at $b_j \in p^{-1}(b)$ is the unique path $\tilde{\omega} : [0, 1] \to \tilde{X}$ such that $\tilde{\omega}(0) = b_j$ and $p \circ \tilde{\omega} = \omega$. Since ω is a loop, the endpoint $\tilde{\omega}(1)$ must be in $p^{-1}(b)$; this gives a natural action of the fundamental group $\pi_1(X, b)$ on the fiber $p^{-1}(b)$.

Definition: Under the *monodromy action* of the fundamental group $\pi_1(X, b)$ on the fiber $p^{-1}(b)$, each $\omega \in \pi_1(X, b)$ acts on the set $p^{-1}(b)$ by sending $b_j \in p^{-1}(b)$ to $\tilde{\omega}_{b_j}(1)$, where $\tilde{\omega}_{b_j}$ is the lift of ω based at b_j. We call the set-automorphism of $p^{-1}(b)$ the monodromy of ω and denote it by $\sigma(\omega)$.

Branched Covering

We will also frequently use the notions of *multigraphs* and *branched coverings* and will therefore define them now:

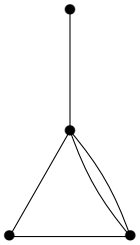
1. By *multigraph* we mean a connected undirected planar graph which allows multiple edges between pairs of vertices but not loops. We shall denote the set of edges of a multigraph H by E and its set of vertices by V.

2. A *branched covering* is an orientation preserving covering map $f : S^2 \to S^2$ which is at least of degree two. For our purposes we shall be interested in branched coverings $f : (S^2 \setminus V) \to (S^2 \setminus V)$ where V is the set of vertices of a multigraph H.

To construct a branched covering from a connected multigraph H, we "blow up" each edge of the given multigraph by cutting along each edge of the graph and opening it up so it is homeomorphic to a closed disk [5]. Next we map the interior of this disk homeomorphically onto the complement of this edge in H in such a way that preserves orientation. The significance of this construction is that each edge in H corresponds to a copy of H in the branched cover, namely the one constructed by blowing up that edge.

To hope to classify critically fixed rational functions, we begin by specifying a degree for f, a number n of critically fixed points, and the multiplicity of each of these critical points. We can represent this data with a multigraph as follows: let H have $\deg(f) - 1$ edges, the same number of vertices as number of critical points, and let each vertex have valence equal to the multiplicity of the critical point it represents. It is possible in general to pick two non-isomorphic graphs to represent this data, and it is known that the resulting graphs yield critically fixed rational functions which do not differ only by Möbius conjugation [4]. Therefore we may think of choosing a multigraph H as narrowing the constraints we place on f [4].

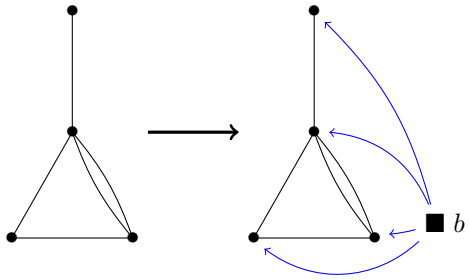
For example, suppose we would like to find a rational function f of degree six, with four critical points of multiplicity 4,3,2, and 1, respectively. We represent this information with the following multigraph:



To find f, we consider H on the sphere $S^2 \cong \hat{\mathbb{C}}$. From this picture we construct the branched covering of H, determine the monodromy of the branched covering, and use the monodromy to find a wreath recursion of $\pi_1(S^2 \setminus V, b)$. Once done, we feed this wreath recursion to Bartholdi's program and it computes a numerical approximation for f.

Our first step is to choose a basepoint b, which we may assume is outside of H. Next we choose a path from b to each vertex V. We insist that

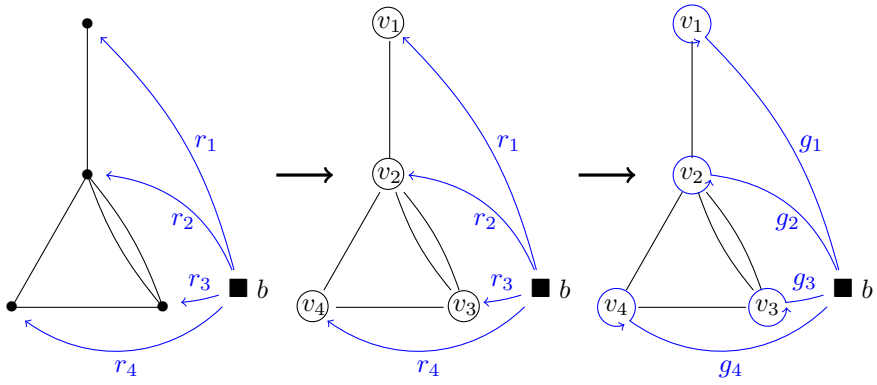
1. These paths not intersect each other except at b,
2. These paths not cross an edge adjacent to the vertex they run to,
3. These paths not cross an edge in E more than once.



Remark: The existence of these paths r_i can be shown by an application of the greedy algorithm on the dual graph of H and a choice of a spanning tree.

We give these paths a cyclic order by labeling r_1 and numbering the indices of the rest in ascending order, moving counter clockwise around the basepoint b for $2 \leq i \leq n = |V|$. We label each $v_i \in V$ with the index of the path r_i which connects it to b.

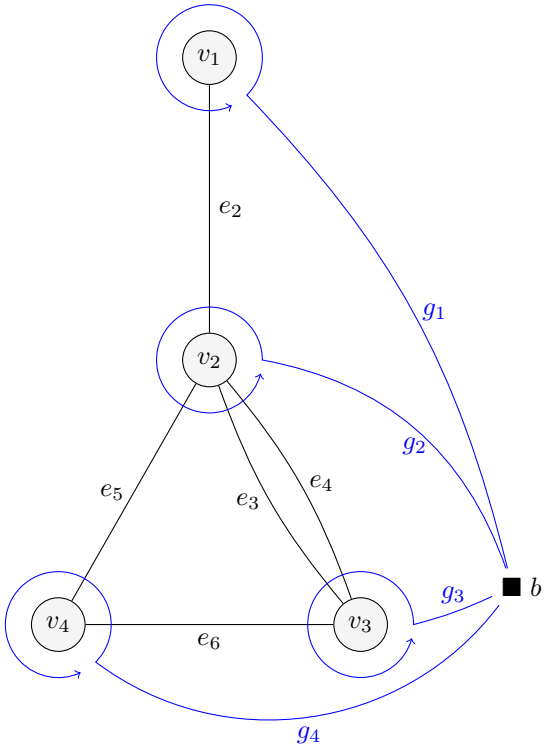
For $1 \leq i \leq n$, let g_i be the loop which follows r_i to within an ϵ-distance of v_i, circles counter clockwise around v_i, and returns to b along r_i.



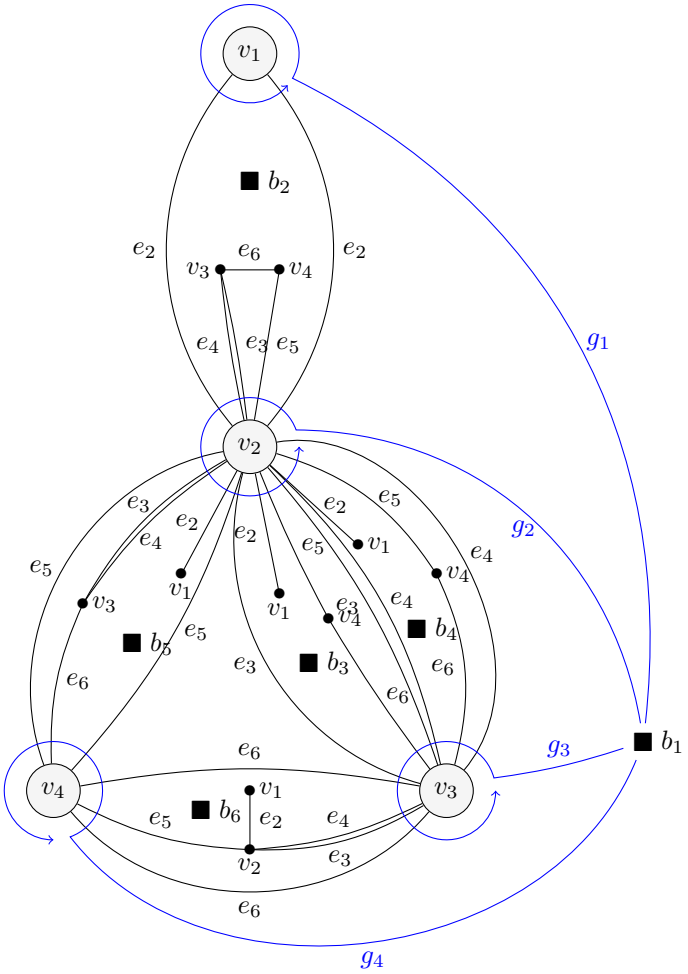
The elements $g_1, \ldots, g_n$ generate $\pi_1(S^2 \setminus V, b)$. Because of this, we shall refer to the r_i's as *generator stems*. Since $\pi_1(S^2 \setminus V, b)$ is the fundamental group of the n-times punctured sphere, it must be isomorphic to the free group on $n-1$ generators. As $g_1, \ldots, g_n$ is a set of n generators, there must be a relation on the g_i. We see it is $g_1 * \cdots * g_n \simeq b$, which arises from the fact that the generators are cyclically ordered. It is also seen readily by stretching $g_1 * \cdots * g_n$ around the back of the sphere.

Now we label the edges of H. Considering parallel edges as distinct, we label the edges in E with the elements in the set $\{e_2, \ldots, e_d\}$ where $d = \deg(f) = |E| + 1$. Beginning with the edges connected to v_1, we label in ascending order the edge which connects v_1 to the vertex with the lowest index. Once all of the edges adjacent to v_1 are labeled, we repeat this process for each v_i connected to v_1 by an edge in E, in order of ascending index. In the case of multiple edges, we number the edges counting upwards as we move counter-clockwise around the vertex in question from the incoming path r_i. We continue until every edge in the graph has been labeled, as shown below.

The significance of this labeling is that it induces a labeling of the set $f^{-1}(b)$, and does so without introducing other choices. By construction, the branched covering f gives a correspondence between edges in E and copies of H in the branched cover that arise from blowing up edges in E. Since each blowup of an edge in E contains exactly one copy of the basepoint, the correspondence is one-to-one: we denote the point in $f^{-1}(b)$ which corresponds to the blow up of e_j by b_j. We label the single element of $f^{-1}(b)$ which does not arise from blowing up an edge b_1.



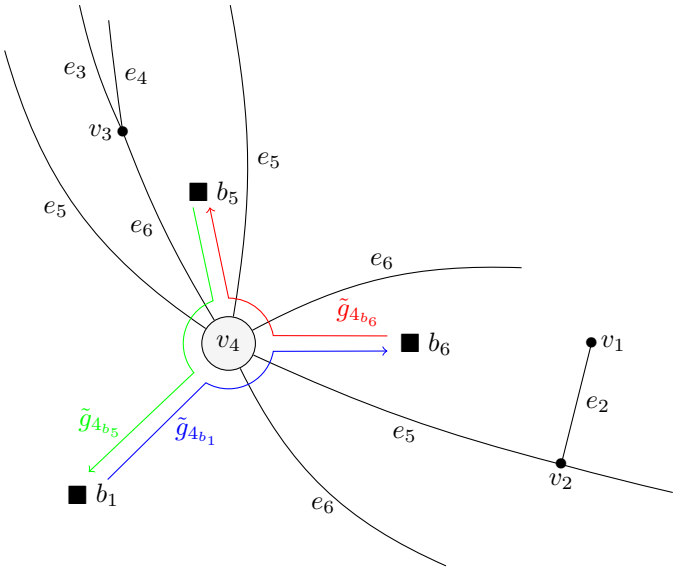
Now we construct the branched covering f. We begin with the identity map $f_{ID} : (S^2 \setminus V) \to (S^2 \setminus V)$. Each edge $e_j \in E$ is an embedded arc in $(S^2 \setminus V)$ on which f_{ID} is injective. We cut along the arc e_j and open the slit slightly to obtain the two arcs $e_{j\pm}$ which together bound a disc D. We send the interior of D to the exterior of $f_{ID}(e_j)$ and the exterior of D first homeomorphically to the complement of e_j, then via f. We send the boundary $e_{j\pm}$ of D to e_j. Repeating this process for each edge $e_j \in E$, we obtain the following branched cover.



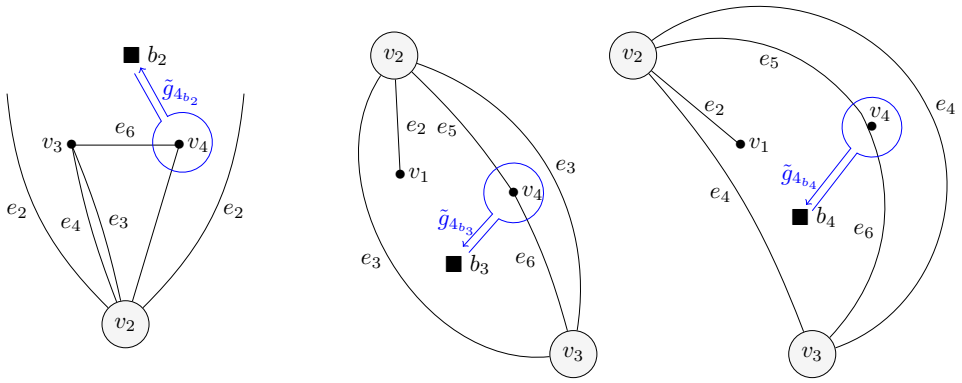
Now we can compute the monodromy action of $\pi_1(S^2 \setminus V, b)$. Recall that under this action $\omega \in \pi_1(S^2 \setminus V, b)$ acts on $f^{-1}(b)$ by sending $b_j \in f^{-1}(b)$ to $\tilde{\omega}_{b_j}(1)$, where $\tilde{\omega}_{b_j}$ is the lift of ω based at $\tilde{b}_j$. As it is a group action, it is completely determined by the action of the generators of $\pi_1(S^2 \setminus V, b)$. In our example, however, rather than compute $\tilde{g}_{i_{b_j}}(1)$ for each $1 \leq i \leq 4$ and $1 \leq j \leq 6$, we shall only compute $\sigma(g_4)$.

We begin with $\tilde{g}_{4_{b_1}}(1)$. Since f is a local homeomorphism, $\tilde{g}_{4_{b_1}}$ can only cross copies of the edges that g_4 crosses in H. Since g_4 runs from b through e_6, through e_5 and returns to b, $\tilde{g}_{4_{b_1}}$ must run from b_1 through a copy of e_6, through a copy of e_5, to an element in $f^{-1}(b)$. Thus $\tilde{g}_{4_{b_1}}(1) = b_6$. As the lifts $\tilde{g}_{4_{b_6}}$ and $\tilde{g}_{4_{b_5}}$ are similar to $\tilde{g}_{4_{b_1}}$, we include them in the following inset of the

region surrounding v_4.



For elements b_j in $f^{-1}(b)$ which correspond to edges e_j not adjacent to v_4, we see that the interior of the blowup of e_j contains a copy of v_i, and thus the lift $\tilde{g}_{4_{b_j}}$ circles this copy of v_i and returns to b_j within the blowup of e_j so that $\tilde{g}_{4_{b_j}}(1) = b_j$.



Thus

$$\tilde{g}_{4_{b_1}}(1) = b_6 \qquad \tilde{g}_{4_{b_2}}(1) = b_2 \qquad \tilde{g}_{4_{b_3}}(1) = b_3$$
$$\tilde{g}_{4_{b_4}}(1) = b_4 \qquad \tilde{g}_{4_{b_5}}(1) = b_1 \qquad \tilde{g}_{4_{b_6}}(1) = b_5$$

so that $\sigma(g_4) = (16)(2)(3)(4)(51)(65) = (165)$.

To determine the monodromy action in full, we must compute $\sigma(g_i)$ for each generator g_i. But since there are d elements in $f^{-1}(b)$, this means computing $n \cdot d$ path lifts in addition to constructing a branched covering. Fortunately, we notice that our careful labeling of H provides a method for computing $\sigma(g_i)$ for each g_i without performing a single path lift, or even constructing a branched covering.

3 Monodromy Algorithm

Step 1

Given a multigraph H, choose a basepoint b and cyclically ordered paths r_i from b to each vertex, for each $1 \leq i \leq n$, such that $r_i \cap r_j = \{b\}$ for all $i \neq j$.

Step 2

Label each vertex v_i with the index of the path r_i that connects it to b. Choose generators g_i to follow each r_i to v_i, loop counter-clockwise around v_i, and return to b along r_i.

Step 3

Label the edges as described above. Beginning with the edges connected to v_1, label in ascending order the edges which connect v_1 to verticies v_k of ascending index k. Once all of the edges adjacent to v_1 are labeled, repeat this process for each v_i connected to v_1 by an edge in E, in order of ascending index i. In the case of multiple edges, number the edges counting upwards, moving counter-clockwise around the vertex in question from the incoming path r_i. Continue until every edge in the graph has been labeled.

Step 4

For each $1 \leq i \leq n$, let the set of edges in E which r_i crosses on its way to v_i divide r_i into path segments. Label each such segment e_k, where e_k is the edge r_i has most recently crossed. If r_i has not yet crossed an edge, label that segment e_1.

Theorem 1: For each $1 \leq i \leq n$, draw a circle C_i of radius ϵ around v_i. Starting with the label on the path r_i, record in order the indices of the labels of the edges which intersect C_i, moving counter-clockwise around C_i from its intersection with r_i. This list is the monodromy $\sigma(g_i)$.

Proof: We divide the proof into two lemmas.

Lemma 1: If neither an edge adjacent to the vertex v_i nor the last segment of r_i is labeled e_j, then the monodromy action of g_i on b_j is trivial.

Proof of Lemma 1: We treat the case $j = 1$ first. If $j = 1$, we wish to show $\tilde{g}_{i_{b_1}}(1) = b_1$. The path r_i must cross an edge in E before it reaches v_i since otherwise the label of the last segment of r_i would be e_1. Additionally, since we assumed r_i crosses no edges adjacent to v_i, it cannot be that v_i is adjacent to the first edge e_k that r_i crosses. Thus the copy of v_i contained in the blowup of e_k must be contained in the interior of the blowup of e_k and therefore have local degree 1. Since e_k is the first edge g_i crosses, $\tilde{g}_{i_{b_1}}$ will travel along $\tilde{r}_{i_{b_1}}$ from b_1 into the disk obtained by blowing up e_k. Since g_i does not cross back through e_k until after circling v_i, $\tilde{g}_{i_{b_1}}$ circles the copy of v_i in the blowup of e_k before returning through e_k to b_1, crossing no more edges on its way. Thus $\tilde{g}_{i_{b_1}}(1) = b_1$.

For $2 \leq j \leq d$ if r_i does not cross the edge e_j on its way to v_i, then $\tilde{g}_{i_{b_j}}$ cannot leave the interior of the blowup of e_j as its boundary is composed of copies of e_j. Since the only element of $f^{-1}(b)$ within the blowup of e_j is b_j, it must be that $\tilde{g}_{i_{b_j}}(1) = b_j$. On the other hand, if r_i crosses e_j on its way to v_i, then v_i cannot be adjacent to e_j. Furthermore, e_j cannot be the last edge that r_i crosses as this implies the last segment of r_i is labeled e_j. Thus $\tilde{g}_{i_{b_j}}$ follows $\tilde{r}_{i_{b_j}}$ through the blowup of e_j into the blowup of e_k, the edge r_i crosses first after crossing e_j. Since e_k is not adjacent to v_i, its blowup will contain a copy of v_i in its interior. Thus $\tilde{g}_{i_{b_j}}(1) = b_j$ by what we showed above. □

Lemma 2: If either an edge adjacent to v_i or the last segment of r_i is labeled e_j, then the generator g_i acts nontrivially on b_j. In fact, g_i acts by sending b_j to $b_{j'}$ where j' is the index of the first edge or segment of r_i adjacent to v_i counter-clockwise of e_j.

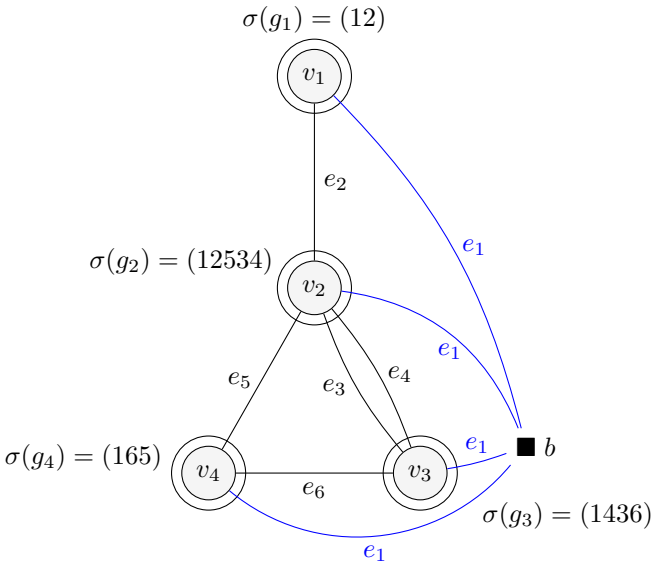
Proof of Lemma 2: Since g_i follows r_i and r_i cannot cross an edge adjacent to v_i, g_i crosses e_j exactly once: namely when g_i circles v_i. This says $\tilde{g}_{i_{b_j}}$ leaves the blowup of e_j and never returns; hence $\tilde{g}_{i_{b_j}} \neq b_j$ and the action must be nontrivial.

If g_i, after crossing e_j, does not cross another edge in E before returning to b, $\tilde{g}_{i_{b_j}}(1) = b_1$. As r_i cannot cross any edges in this case, it must consist of exactly one segment labeled e_1. This agrees with our claim since in this case r_i must be the first edge or segment adjacent to v_i counter-clockwise of e_j.

On the other hand, suppose $e_{j'}$ is the first edge g_i crosses after crossing e_j. Then $e_{j'}$ is either adjacent to v_i or not. If $e_{j'}$ is adjacent to v_i, then the edge $e_{j'}$ is the first edge or segment counter-clockwise of e_j. That $e_{j'}$ is adjacent to v_i also implies r_i cannot cross $e_{j'}$ and therefore g_i must cross $e_{j'}$ exactly once. Thus once $\tilde{g}_{i_{b_j}}$ enters the blowup of $e_{j'}$ it cannot leave. Since the only element of $f^{-1}(b)$ in the blowup of $e_{j'}$ is $b_{j'}$, we conclude $\tilde{g}_{i_{b_j}}(1) = b_{j'}$.

Finally, suppose $e_{j'}$ is the first edge g_i crosses after crossing e_j, but that $e_{j'}$ is not adjacent to v_i. Then the segment of r_i adjacent to v_i is the first edge or segment of r_i adjacent to v_i counter-clockwise of e_j. This also implies $e_{j'}$ is the last edge r_i crosses before reaching v_i and therefore that the segment of r_i adjacent to v_i is labeled $e_{j'}$. Therefore we need only show $\tilde{g}_{i_{b_j}} = b_{j'}$. But this is clear since r_i can cross an edge no more than once: that is, once $\tilde{g}_{i_{b_j}}$ enters the blowup of $e_{j'}$ it cannot leave. Then $\tilde{g}_{i_{b_j}} = b_{j'}$. □

This says the monodromy action $\sigma(g_i)$ of each generator of $\pi_1(S^2 \setminus V, b)$ consists of exactly one cycle, and that this cycle consists of the indices of the labels of the edges and segment of r_i which is adjacent to v_i, written moving counter-clockwise around v_i. The practicality of this method is illustrated in computing the monodromy of our example, which can be read directly off of the graph.



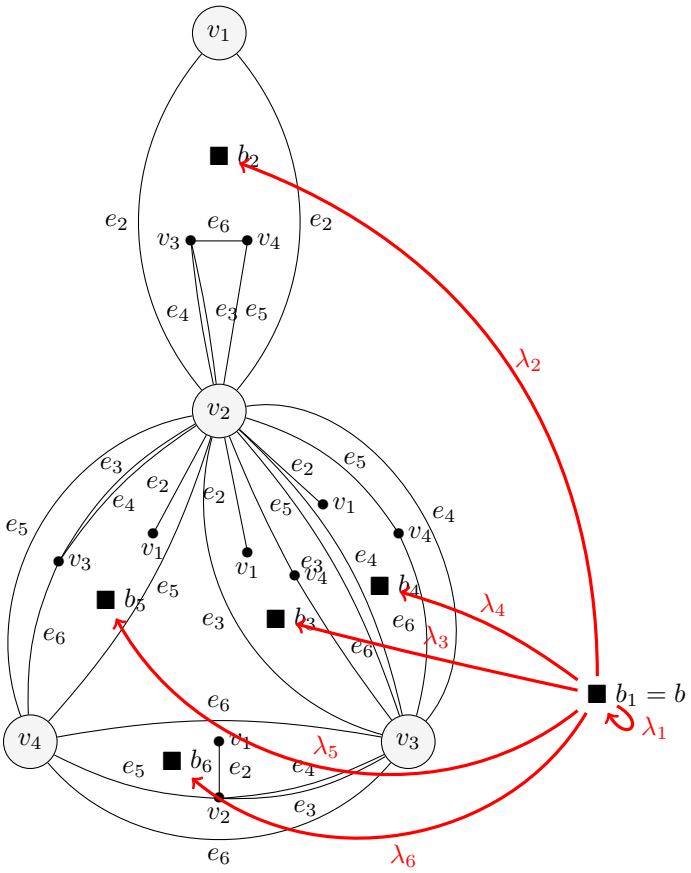
Remark: As it will be convenient later, we insist that each cycle $\sigma(g_i)$ begin with the index of the label on the segment of r_i adjacent to v_i. This will allow us to refer to the "first" and "last" number in the monodromy $\sigma(g_i)$.

4 Wreath Recursions

The *wreath recursion* of a group G is a homomorphism $\Phi : G \longrightarrow G^d \rtimes S_d$ from G to the semi-direct product of G^d and S_d, where S_d acts on G^d by permuting its coordinates. In what follows we will give substance to this definition in the case where $G = \pi_1(S^2 \setminus V, b)$ and explain an effective method for computing Φ by

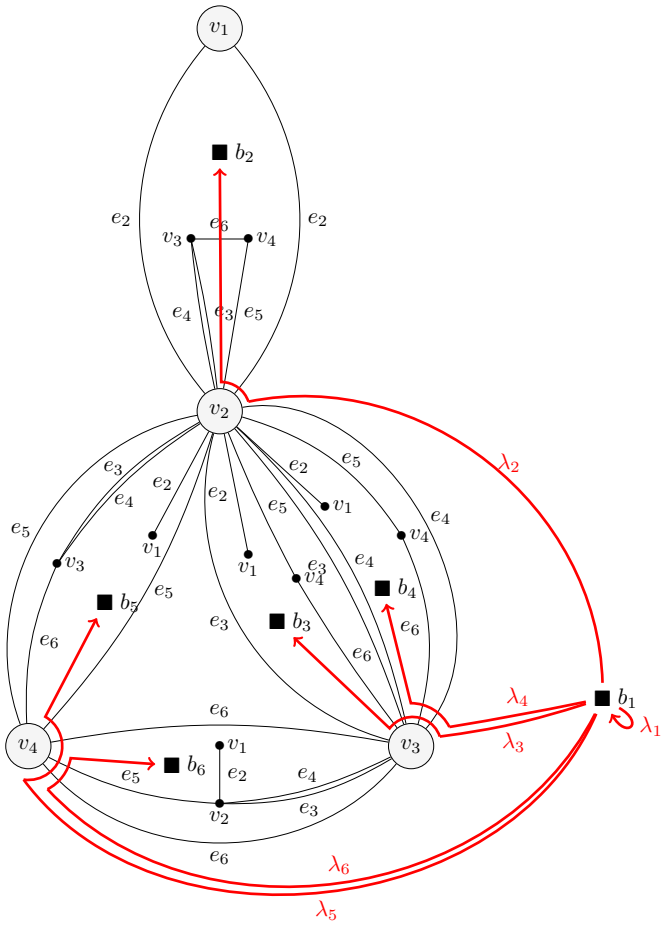
hand. Afterwards we shall describe an algorithm to find Φ, given a multigraph H and a choice of generators.

We must first introduce the notion of a connecting path. Already we have chosen a basepoint b, a labeling $\{1, \ldots, d\}$ of the elements of $f^{-1}(b)$, and generators $g_1, \ldots, g_n$ which run from b to each $v_i \in V$, loop counter-clockwise around v_i, and return to b. Now for each $j \in \{1, \ldots, d\}$ we choose a *connecting path* $\lambda_j : [0,1] \to S^2 \setminus V$ running from the basepoint b to the element b_j of $f^{-1}(b)$ corresponding to j. As b and b_1 are in the same region, we identify them. The connecting paths are shown in red.

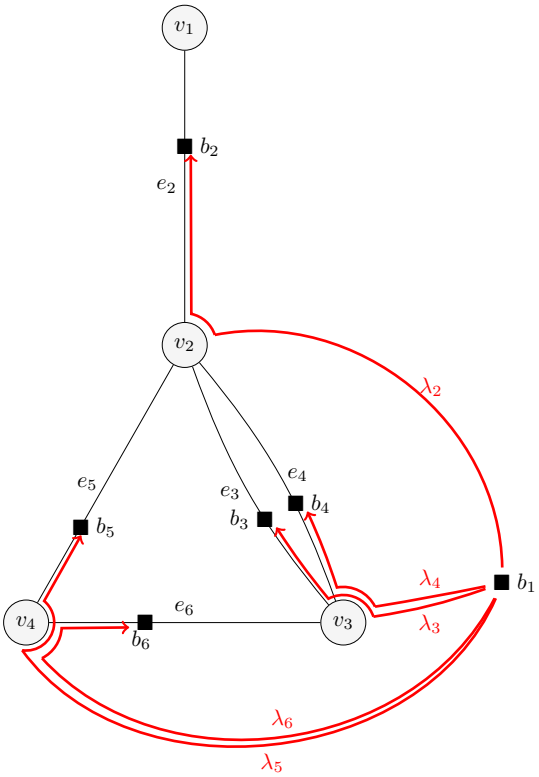


We wish to find an algorithmic method for choosing connecting paths λ_j from b to b_j, for $1 \leq j \leq d$. Since we have identified b and b_1, we let λ_1 be the constant path at b_1. For $2 \leq j \leq d$, we recall that we have identified $b_j \in f^{-1}(b) \setminus b_1$ with the edge $e_j \in E$. Thus for $j > 1$, rather than thinking of λ_j as a path from b to b_j, we may think of it as a path from b to the edge e_j.

As we forbid H to contain loops, there are exactly two distinct vertices v_k, v_l adjacent to e_j. Since we've chosen g_k to circle v_k and g_l to circle v_l, we already have two paths from b to e_j: the first follows g_k until it intersects e_j, the second does the same but follows g_l instead of g_k. Since it is always the case that if $k \neq l$ either $k > l$ or $l > k$, we may always choose λ_j so that if $k > l$, λ_j is the path which follows g_k until it enters the blowup of the edge e_j and then follows any path to b_j that is contained entirely within the interior of the blowup of e_j. Since the interior of the blowup of e_j contains no points in V, any two paths contained entirely within the blowup of e_j which have matching startpoints and endpoints are homotopic.



Since λ_j may take any path to b_j once it has entered the blowup, we do not lose any generality by drawing the connecting paths on the target graph H rather than on its branched covering.



Now we give the definition of a wreath recursion in its entirety.

Definition: Suppose H is a planar multigraph with vertex set V, and $f : (S^2 \backslash V) \to (S^2 \backslash V)$ is a degree d branched covering constructed from H. Suppose further that $b \in S^2 \setminus V$, $f^{-1}(b) \leftrightarrow \{1, \ldots, d\}$ is a labeling of $f^{-1}(b)$, $g_1, \ldots, g_n$ are generators of $\pi_1(S^2 \setminus V, b)$, and $\lambda_j : [0,1] \to S^2 \setminus V$ is a connecting path from b to b_j for every $1 \leq j \leq d$. We say the *wreath recursion* of $G = \pi_1(S^2 \setminus V, b)$ is a homomorphism

$$\Phi : G \longrightarrow (\underbrace{G \times \cdots \times G}_{d-\text{times}}) \rtimes S_d = G^d \rtimes S_d \text{ which sends } g \mapsto \langle g|_1, \ldots, g|_d \rangle \sigma(g).$$

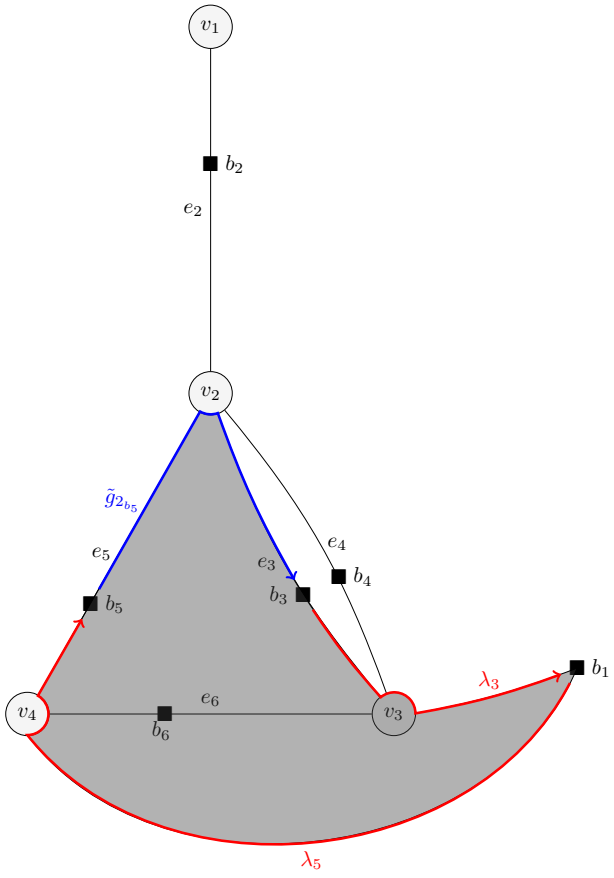
where $g|_j = [\lambda_j * \tilde{g}_{b_{j^{\sigma(g)}}} * \overline{\lambda}_{j^{\sigma(g)}}]$. By convention, we write $j^{\sigma(g_i)}$ for $j \cdot \sigma(g_i)$, the right action of $\sigma(g_i)$ on j.

The symbol $\rtimes$ denotes the semidirect product by which S^d acts on G^d by permuting its coordinates. Multiplication in $G^d \rtimes S_d$ is defined as follows: $(\langle g_1, \ldots, g_d \rangle \sigma) * (\langle h_1, \ldots, h_d \rangle \tau) = (g_1 h_{1^\sigma}, \ldots, g_d h_{d^\sigma}) \sigma\tau$.

5 Computing a Wreath Recursion

To provide some intuition for this process we shall compute a wreath recursion Φ of our example $\pi_1(S^2 \setminus V, b)$. Since Φ is a homomorphism, it is enough to compute $\Phi(g_i)$ for each generator g_i of $\pi_1(S \setminus V, b)$. For the sake of brevity we shall only give the computation of the wreath position $g_2|_5$ and informally describe several helpful lemmas which we shall state and prove explicitly in the proof of our general algorithm.

To determine $g_2|_5 = [\lambda_5 * \tilde{g}_{2_{b_5}} * \overline{\lambda}_{5^{\sigma(g_2)}}]$, we must first find λ_5. Since the edge e_5 is connected to the vertices v_2 and v_4, and $4 > 2$, λ_5 follows the generator g_4 from b until it intersects the edge e_5. Then it follows e_5 until it reaches b_5. The monodromy $\sigma(g_2) = (12534)$ tells us that the lift $\tilde{g}_{2_{b_5}}$ is homotopic to the path which follows edge e_5 from b_5 until it intersects g_2, follows g_2 counter-clockwise around v_2 until it intersects $e_{5^{\sigma(g_2)}} = e_3$, and follows e_3 to b_3. Since e_3 is connected to vertices v_2 and v_3, and $3 > 2$, $\overline{\lambda}_{j^{\sigma(g_2)}} = \overline{\lambda}_3$ follows g_3 clockwise around v_3 from where it intersects e_3 back to the basepoint b.



By examining the region this path $g_2|_5$ encloses (shaded in the picture), we see that $\lambda_5 * \tilde{g}_{2_{b_5}} * \overline{\lambda}_{5^{\sigma(g_2)}}$ is homotopic to g_3^{-1} and thus conclude $g_2|_5 = g_3^{-1}$.

Performing this computation for each wreath position $g_i|_j$, $1 \leq i \leq 4, 1 \leq j \leq 6$, we notice several striking patterns. First, if j is fixed by the monodromy $\sigma(g_i)$ then $g_i|_j$ is trivial. We delay the rigorous statement and proof of this fact, but the intuition is latent in the procedure we have described: if v_k is the vertex of higher degree attached to the edge e_j, then λ_j follows g_k to e_j. Since j is fixed by the monodromy, $\tilde{g}_{i_{b_j}}$ stays within the disk obtained by blowing up e_j and thus $\overline{\lambda}_{j^{\sigma(g_i)}} = \overline{\lambda}_j$ traverses λ_j backwards, yielding a path homotopic to the constant map at b. Together with our computation $g_2|_5 = g_3^{-1}$, this gives us 11 of the 24 wreath positions:

$$\begin{aligned}
\Phi(g_1) &= \langle\ ,\ ,\ e,\ e,\ e,\ e\rangle(12)\\
\Phi(g_2) &= \langle\ ,\ ,\ ,\ ,\ g_3^{-1},\ e\rangle(12534)\\
\Phi(g_3) &= \langle\ ,\ e,\ ,\ ,\ e,\ \rangle(1436)\\
\Phi(g_4) &= \langle\ ,\ e,\ e,\ e,\ ,\ \rangle(165)
\end{aligned}$$

The second observation we make is that if v_i is the vertex with higher index attached to both the edge e_j and the edge $e_{j^{\sigma(g_i)}}$, then the wreath position $g_i|_j$ is trivial. In this case λ_i follows g_i to the edge e_j and runs up e_j to b_j. Then the lift $\tilde{g}_{i_{b_j}}$ travels back down e_j to g_i, follows g_i to $e_{j^{\sigma(g_i)}}$ and runs up $e_{j^{\sigma(g_i)}}$ to $b_{j^{\sigma(g_i)}}$. Then since v_i is the vertex of higher index attached to $e_{j^{\sigma(g_i)}}$, $\overline{\lambda}_{j^{\sigma(g_i)}}$ goes back down $e_{j^{\sigma(g_i)}}$ to g_i and follows g_i out backwards to b. Thus $g_i|_j$ encircles no elements of V and is therefore homotopic to the constant map. This gives us five more wreath positions.

$$\begin{aligned}
\Phi(g_1) &= \langle\ ,\ ,\ e,\ e,\ e,\ e\rangle(12)\\
\Phi(g_2) &= \langle\ e,\ ,\ ,\ ,\ g_3^{-1},\ e\rangle(12534)\\
\Phi(g_3) &= \langle\ e,\ e,\ ,\ e,\ e,\ \rangle(1436)\\
\Phi(g_4) &= \langle\ e,\ e,\ e,\ e,\ ,\ e\rangle(165)
\end{aligned}$$

Now we fill in the rest of the table, encouraging the reader to check our work using the processes described above.

$$\begin{aligned}
\Phi(g_1) &= \langle\ g_1,\ e,\ e,\ e,\ e,\ e\rangle(12)\\
\Phi(g_2) &= \langle\ e,\ g_2g_3,\ e,\ e,\ g_3^{-1},\ e\rangle(12534)\\
\Phi(g_3) &= \langle\ e,\ e,\ g_3,\ e,\ e,\ e\rangle(1436)\\
\Phi(g_4) &= \langle\ e,\ e,\ e,\ e,\ g_4,\ e\rangle(165)
\end{aligned}$$

Recall the relation $g_1g_2g_3g_4 = e$ on the generators and the fact that Φ is a homomorphism. Since $\Phi(g_1g_2g_3g_4) = \Phi(g_1)\Phi(g_2)\Phi(g_3)\Phi(g_4)$, it must be that

$\Phi(g_1)\Phi(g_2)\Phi(g_3)\Phi(g_4) = e$. This requires some familiarity with multiplication on semidirect products, but it can be an efficient way to check one's work.

6 Wreath Recursion Algorithm

Our goal is to express each wreath position $g_i|_j$ in terms of the generators $g_1, \ldots, g_n$ of $\pi_1(S^2 \setminus V, b)$. We shall assume we have performed the monodromy algorithm described above and that we have the following input.

Input: For each edge $e_j \in E$, we input the list of generators of $\pi_1(S^2 \setminus V, b)$ which intersect e_j, ordered from the generator which intersects e_j nearest the vertex adjacent to e_j with the lower index to the generator which intersects e_j nearest the vertex adjacent to e_j with the higher index.

Our guiding principle will be a procedure we describe at the end of the paper that writes any loop $\omega \in \pi_1(S^2 \setminus V, b)$ as a word in the generators g_i, provided we know the following data:

1. The generators g_s, g_{s+1} that ω starts "between" and the generators g_e, g_{e+1} that ω ends "between,"
2. The ordered list $c_1, \ldots, c_k$ of the generators that ω intersects away from b, written in the order ω intersects them,
3. The orientation, positive or negative, with which ω intersects the generator corresponding to c_x, for $1 \leq x \leq k$.

Remark: Since we may ϵ-perturb the g_i's in a continuous manner, we may assume ω starts and ends "between" two generators g_i, since, after ϵ-perturbation, we may choose some $\delta > 0$ such that $(\cup_{i=1}^{n} g_i) \bigcap \omega = \{b\}$ within a δ-ball centered at b. Furthermore, the orientation of the intersection of ω with the generator corresponding to c_x is well defined since we may ϵ-perturb each g_i so its intersection with ω is transverse.

We say ω intersects the generator g_i with *positive orientation* if a person walking along ω would find the basepoint connected to the segment of g_i on the person's left at the point of intersection. Otherwise, we say ω intersects g_i with *negative orientation*. In light of this procedure, our goal of expressing $g_i|_j$ in terms of the g_i reduces to finding the data (1), (2), and (3).

As we have observed, each $g_i|_j$ is homotopic to a path which follows only segments of generators g_i and segments of edges $e_j \in E$. What is more, we have described a process for finding this representative of $g_i|_j$ based on a small subset of the data in the pictures we draw: namely the edges in E and segments of generator stems r_i labeled e_j and $e_{j^{\sigma(g_i)}}$, their adjacent vertices, and the generators which circle these vertices.

Definition: We call the picture composed of the edges in E and the segments of generator stems labeled e_j and $e_{j^{\sigma(g_i)}}$ along with their adjacent vertices and the generators that circle them the *local picture* of $g_i|_j$.

Lemma 3: Away from the basepoint b, the loop $g_i|_j$ can only intersect generators of $\pi_1(S^2 \setminus V, b)$ that cross the edges e_j or $e_{j^{\sigma(g_i)}}$ in E.

Proof: This follows directly from the fact that $g_i|_j$ is homotopic to a path that follows only generators and the edges $e_j, e_{j^{\sigma(g_i)}} \in E$. □

To find $c_1, \ldots, c_k$, we will determine the segments of e_j and $e_{j^{\sigma(g_i)}}$ that $g_i|_j$ traverses and consult the lists of the generators that cross e_j and $e_{j^{\sigma(g_i)}}$ that we have inputted. Then it will remain only to find the generators $g_s, g_{s+1}, g_e, g_{e+1}$ that $g_i|_j$ starts and ends between, and the orientation of the intersection of $g_i|_j$ with the generator corresponding to c_x, for all $1 \leq x \leq k$. This data will come from examining the local pictures, of which there are only finitely many.

Thus we begin by classifying the local pictures. Afterwards we will discuss a method for identifying the local picture associated with an arbitrary wreath position $g_i|_j$, describe in detail how to find the data (1), (2), and (3) for $g_i|_j$ from this local picture, and introduce the procedure which uses (1), (2), and (3) to write $g_i|_j$ in terms of the generators $g_1, \ldots, g_n$.

The Classification of Local Pictures

Lemma 4: Suppose $\sigma(g_i)$ acts trivially on j. Then $g_i|_j$ is trivial.

Proof: We wish to show $\lambda_j * \tilde{g}_{i_{b_j}} * \overline{\lambda}_{j^{\sigma(g_i)}} \simeq b$. That $\sigma(g_i)$ acts trivially on j, or, in other words, that $j^{\sigma(g_i)} = j$ implies that $\tilde{g}_{i_{b_j}}$ is a loop as $\tilde{g}_{i_{b_j}}(1) = b_{j^{\sigma(g_i)}} = b_j$. It means also that $\overline{\lambda}_{j^{\sigma(g_i)}} = \overline{\lambda}_j$. Thus we need only check that the loop $\tilde{g}_{i_{b_j}}$ is contractible, or homotopic to the constant path. But this is clear since ω is a loop freely homotopic to a loop surrounding $\tilde{v}_i \in f^{-1}(v_i)$, where $\tilde{v}_i \neq v_i$. Thus ω is contractible in $S^2 \setminus V$.

Explicitly, by construction we may choose $\epsilon > 0$ so that g_i follows r_i from b to within the ϵ-ball B centered at $v_i \in V$, circles v_i within B, and returns to b along r_i. Denote the copy of v_i in the blowup of e_j by $\tilde{v}_i$. Since v_i is not adjacent to e_j, $\tilde{v}_i$ must be in the interior of the blowup of e_j. Now choose a δ-ball B' about $\tilde{v}_i$ in the interior of the blowup of e_j such that B is contained in the image of B' under f. Since we constructed the branched covering to send the interior of the blowup of an edge homeomorphically to the complement of that edge in $S^2 \setminus V$, f is a homeomorphism when restricted to B'. Therefore $\tilde{g}_{i_{b_j}}$ follows $\tilde{r}_{i_{b_j}}$ into B', circles $\tilde{v}_i$ within B', and returns to b_j along $\tilde{r}_{i_{b_j}}$. As $\tilde{v}_i \notin V$, $\tilde{g}_{i_{b_j}}$ is contractible. □

By Theorem 1 we have that $\sigma(g_i)$ acts trivially on j if and only if neither an edge in E adjacent to v_i nor the segment of r_i adjacent to v_i is labeled e_j. Having now proven that $g_i|_j$ is trivial in this case, we turn our attention to the cases where either an edge in E adjacent to v_i or the segment of r_i adjacent to v_i is labeled e_j.

We first treat the case where the label of the segment of r_i adjacent to v_i is neither e_j nor $e_{j^{\sigma(g_i)}}$. Since we have insisted the first number in the monodromy be the index of the label of the segment of r_i adjacent to v_i, we may phrase this in terms of the position of j in $\sigma(g_i)$.

Case 1 : j is neither the first nor the last number in the monodromy $\sigma(g_i)$

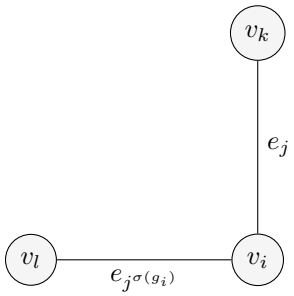
Any $g_i|_j$ specifies a vertex v_i and an edge or generator segment labeled e_j. If we assume that j is neither the first nor the last number in the monodromy $\sigma(g_i)$, then e_j is an edge in E and must be connected to a vertex v_k distinct from v_i. In addition, $e_{j^{\sigma(g_i)}}$ is an edge in E and must be connected to the vertex v_i and a vertex v_l distinct from v_i. That is, if j is neither the first nor the last number in $\sigma(g_i)$, $g_i|_j$ specifies the edges e_j, $e_{j^{\sigma(g_i)}} \in E$ and the vertices $v_i, v_k, v_l \in V$.

Case 1 (a) : v_k and v_l are distinct

If we assume v_k and v_l are distinct we see there are the following six sub-cases:

(1) $k < l < i$	(2) $l < k < i$	(3) $k < i < l$
(4) $l < i < k$	(5) $i < k < l$	(6) $i < l < k$

where the local picture is



■ b

Lemma 5: Suppose $\sigma(g_i)$ acts nontrivially on j, and that j is neither the first nor the last number in the monodromy. Suppose further that $l, k < i$ where v_k and v_l are the vertices attached to the edges e_j and $e_{j^{\sigma(g_i)}}$, respectively, which

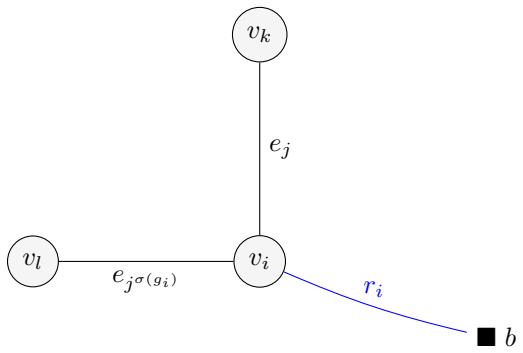
are not v_i. Then $g_i|_j$ is trivial.

Proof: Since $i > k$, λ_i follows g_i until it reaches the blowup of e_j, whereupon it follows any path contained in the interior of the blowup of e_j to b_j. Next, as r_i crosses no edges adjacent to v_i, g_i crosses e_j and $e_{j^{\sigma(g_i)}}$ exactly once: namely when it circles v_i. Therefore $\tilde{g}_{i_{b_j}}$ travels from b_j through the interior of the blowup of e_j, through e_j, then immediately into the blowup of $e_{j^{\sigma(g_i)}}$ and on to b_j. Since $i > l$, $\overline{\lambda}_{j^{\sigma(g_i)}}$ follows any path in the interior of the blowup of $e_{j^{\sigma(g_i)}}$ to the generator g_i and follows $\overline{g}_i$ back out to b. Since there are no elements in V in the interior of the blowup of e_j or $e_{j^{\sigma(g_i)}}$ and v_i is not encircled by $g_i|_j$, the only region in which $g_i|_j$ can encircle a vertex $v \in V$ is the region between the blowup of e_j and the blowup of $e_{j^{\sigma(g_i)}}$. But since $\tilde{g}_{i_{b_j}}$ can cross no edges between its intersections with e_j and $e_{j^{\sigma(g_i)}}$, for $g_i|_j$ to encircle v means v must be connected to v_i via an edge that intersects none of $e_j, e_{j^{\sigma(g_i)}}, \tilde{g}_{i_{b_j}}$. This is impossible as $e_{j^{\sigma(g_i)}}$ is the next edge attached to v_i counter-clockwise of e_j. □

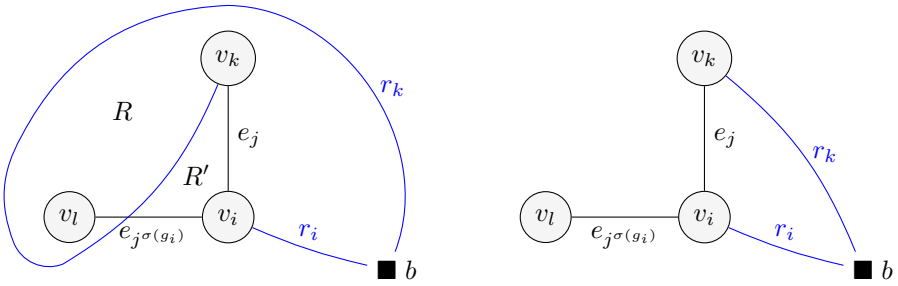
Corollary: Sub-cases (1) and (2) are trivial. □

Sub-case 3: $k < i < l$

Since v_i is adjacent to e_j and $e_{j^{\sigma(g_i)}}$, it cannot be that r_i crosses either e_j or $e_{j^{\sigma(g_i)}}$. Thus there is exactly one homotopy class of paths $[r_i]$ from b to v_i; we choose the path r_i, pictured below, to represent it.

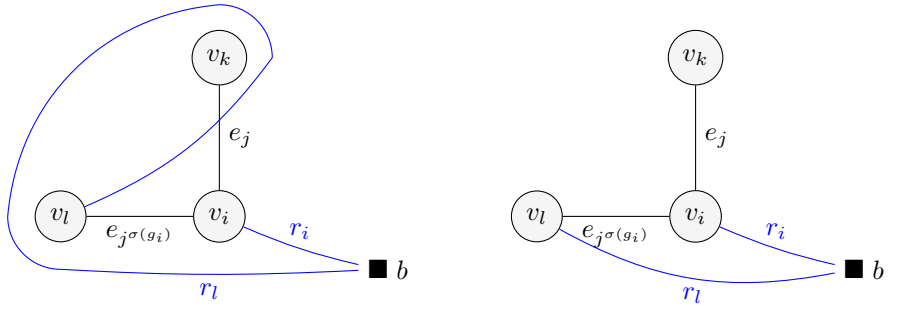


Since v_k is adjacent to e_j, r_k cannot cross e_j. Therefore, as it is still free to cross the edge $e_{j^{\sigma(g_i)}}$, there are exactly two choices for r_k: namely the homotopy class of paths from b to v_k that crosses $e_{j^{\sigma(g_i)}}$ and the homotopy class of paths from b to v_k which does not cross $e_{j^{\sigma(g_i)}}$. We choose the following representatives for these homotopy classes of paths:

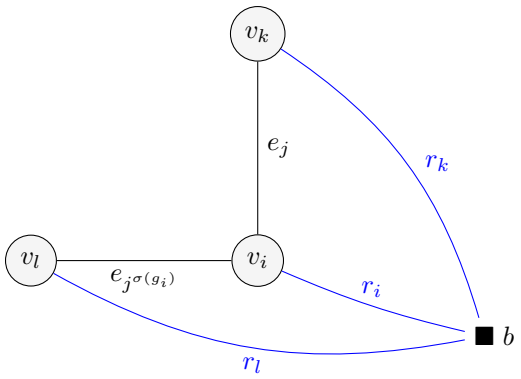


In the case where r_k crosses $e_{j^{\sigma(g_i)}}$, the point v_l is contained in a region R bounded by r_k, r_i, and e_j. Because the generator stems are cyclically ordered and $l > i$, r_l must leave b counter-clockwise of r_i. Thus to enter the region R, r_l must cross e_j since, as distinct generator stems meet only at b, r_l cannot cross r_k or r_i. But to cross e_j, r_k must enter the region R' bounded by r_k, e_j, and $e_{j^{\sigma(g_i)}}$. As r_l cannot intersect r_k, it must cross $e_{j^{\sigma(g_i)}}$, which is a contradiction since v_l is adjacent to $e_{j^{\sigma(g_i)}}$ and r_l is therefore forbidden to cross $e_{j^{\sigma(g_i)}}$. We conclude that r_k does not cross $e_{j^{\sigma(g_i)}}$.

There are also exactly two choices for r_l: the homotopy class of paths from b to v_l that crosses e_j and the homotopy class of paths that does not. We draw each below:



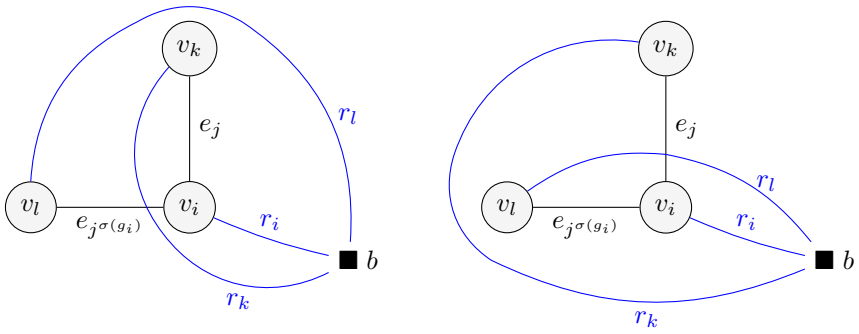
In the case where r_l crosses e_j, v_k is contained in a region bounded by r_l, r_i, and $e_{j^{\sigma(g_i)}}$. Because the generator stems are cyclically ordered and $k < i$, r_k must leave b clockwise of r_i. To enter this region r_k must cross $e_{j^{\sigma(g_i)}}$, which we showed is impossible above. Thus the only choices we have for r_k, r_i, and r_l are the homotopy classes of paths from b to v_k, v_i, and v_l, respectively, which do not cross e_j or $e_{j^{\sigma(g_i)}}$.



Applying similar reasoning we find the following for the remaining three sub-cases of Case 1(a).

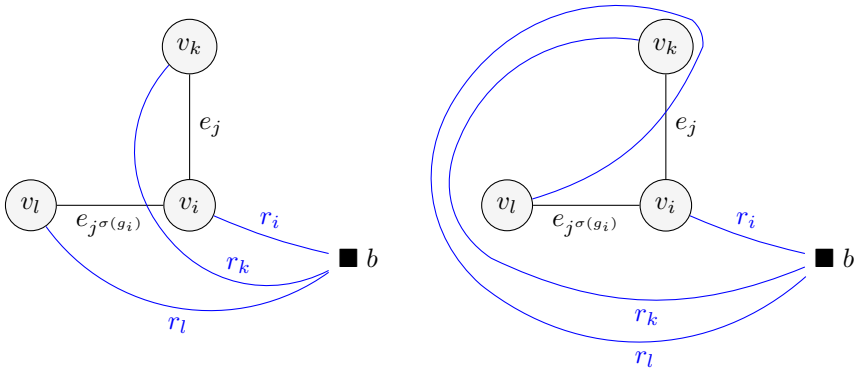
Sub-case 4: $l < i < k$

Either (1) r_k is the homotopy class of paths from b to v_k which crosses $e_{j^{\sigma(g_i)}}$ and r_l is the homotopy class of paths from b to r_l which does not cross e_j or (2) r_k is the homotopy class of paths from b to v_k which does not cross $e_{j^{\sigma(g_i)}}$ and r_l is the homotopy class of paths from b to v_l which does cross e_j.



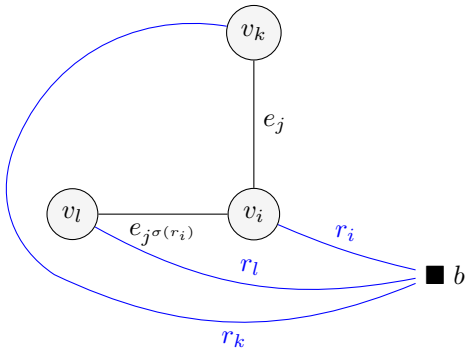
Sub-case 5: $i < k < l$

Either r_k is the homotopy class of paths from b to r_k which crosses $e_{j^{\sigma(g_i)}}$ and r_l is the homotopy class of paths from b to v_l which does not cross e_j, or r_k is the homotopy class of paths from b to v_k which does not cross $e_{j^{\sigma(g_i)}}$ and r_l is the homotopy class of paths from b to v_l which does cross e_j.

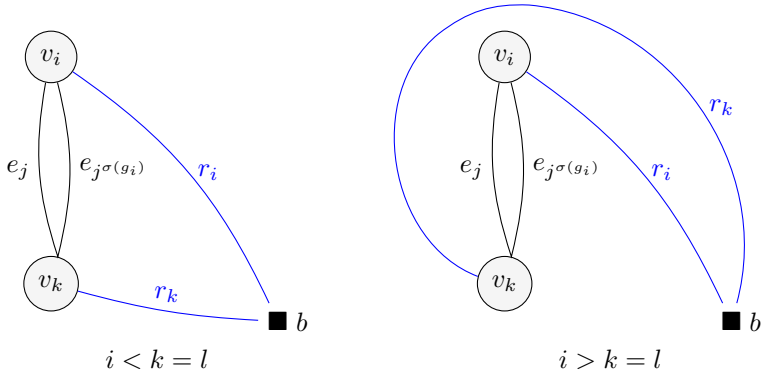


Sub-case 6: $i < l < k$

In this case r_k can be only the homotopy class of paths from b to v_k which does not cross $e_{j^{\sigma(g_i)}}$ and r_l can be only the homotopy class of paths from b to r_l which does not cross e_j.



Case 1 (b): $v_k = v_l$



Lemma 6: If $v_k = v_l$, then $g_i|_j$ is trivial.

Proof. If $i > k = l$, then $g_i|_j$ is trivial by Lemma 5. So suppose $i < k = l$. Let B be the region between the blowups of e_j and $e_{j^{\sigma(g_i)}}$. Let D be the union of B and these blowups. Since e_j and $e_{j^{\sigma(g_i)}}$ form a multiple edge, there can be no element $v \in V$ contained in B. As there can be no $v \in V$ contained in the interior of the blowup of either e_j or $e_{j^{\sigma(g_i)}}$, there can be no element of v contained in D. But $g_i|_j$ is homotopic to a path contained in the interior of D, so $g_i|_j$ must be homotopic to the constant path.

Since e_j and $e_{j^{\sigma(g_i)}}$ are adjacent to v_i, r_i cannot cross either. This says g_i crosses e_j and $e_{j^{\sigma(g_i)}}$ exactly once each: namely when g_i circles v_i. Therefore the lift $\tilde{g}_{i_{b_j}}$ crosses the boundary of the blowup of e_j exactly once and after doing so proceeds directly into the blowup of $e_{j^{\sigma(g_i)}}$, since $e_{j^{\sigma(g_i)}}$ is the edge g_i crosses first after crossing e_j. Since g_i crosses $e_{j^{\sigma(g_i)}}$ exactly once, $\tilde{g}_{i_{b_j}}$ remains within he blowup of $e_{j^{\sigma(g_i)}}$ until it reaches $b_{j^{\sigma(g_i)}}$.

As $k > i$, λ_j and $\lambda_{j^{\sigma(g_i)}}$ each follow g_k into the blowup of $e_{j^{\sigma(g_i)}}$. We may pick some b' in the blowup of $e_{j^{\sigma(g_i)}}$ such that λ_j and $\lambda_{j^{\sigma(g_i)}}$ each pass through b'. Contracting $g_i|_j$ along g_k we see that $g_i|_k$ is homotopic to the loop based at b' which otherwise follows $\lambda_j * \tilde{g}_{i_{b_j}} * \overline{\lambda}_{j^{\sigma(g_i)}}$. Since this loop is contained entirely within D, it must be that $g_i|_j$ is trivial. □

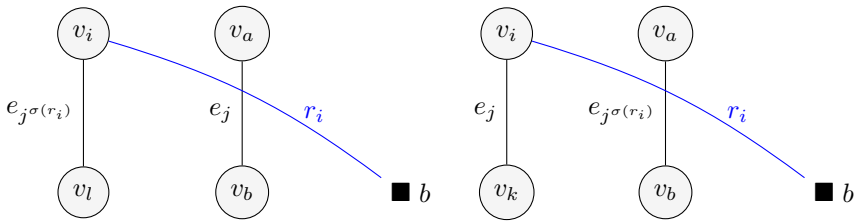
Case 2: j is either the first or last number in the monodromy

In this case it is possible that the edges e_j and $e_{j^{\sigma(g_i)}}$ do not share a common adjacent vertex, and hence it is possible a local picture contains as many as four distinct vertices. To account for this, we let v_a and v_b be the vertices adjacent

to the edge in the local picture which is not adjacent to v_i, if one exists. The vertices v_k and v_l, when they appear, are defined as before.

Case 2(a): $j, j^{\sigma(g_i)} \neq 1;\ v_l \neq v_b, v_k \neq v_b$

Although we have assumed that there are edges in E or segments of generator stems in the local picture labeled e_j and $e_{j^{\sigma(g_i)}}$ adjacent to v_i, we must exercise some caution now as these objects are no longer necessarily edges in E. As $j, j^{\sigma(g_i)} \neq 1$, r_i must cross at least one edge on its way to v_i. If j is the first number in the monodromy, the last edge r_i crosses is e_j. If $j^{\sigma(g_i)}$ is the first number in the monodromy (or, equivalently, if j is the last number in the monodromy), the last edge r_i crosses is $e_{j^{\sigma(g_i)}}$. Recall that the segment of r_i adjacent to v_i is labeled e_k, where e_k is the last edge r_i crosses.



j is the first number in $\sigma(g_i)$ $\quad$ j is the last number in $\sigma(g_i)$

In the case where j is the first number in the monodromy $\sigma(g_i)$, pictured above on the left, $g_i|_j$ follows $g_{\max\{a,b\}}$ to e_j, goes along e_j to g_i, along g_i to $e_{j^{\sigma(g_i)}}$, then along $e_{j^{\sigma(g_i)}}$ to $v_{\max\{i,l\}}$ and back to b along $\overline{g}_{\max\{i,l\}}$. In the case where j is the last number in the monodromy, $g_i|_j$ follows the same path in the opposite direction, except that it still follows g_i counter-clockwise around v_i.

A priori, there are 48 sub-cases of Case 2(a): 4! ways of ordering the set $\{i, k, a, b\}$ and another 4! ways of ordering $\{i, l, a, b\}$.

Lemma 7: Suppose the conditions of Case 2(a) are met. Then permuting the labels a and b has no effect on $g_i|_j$.

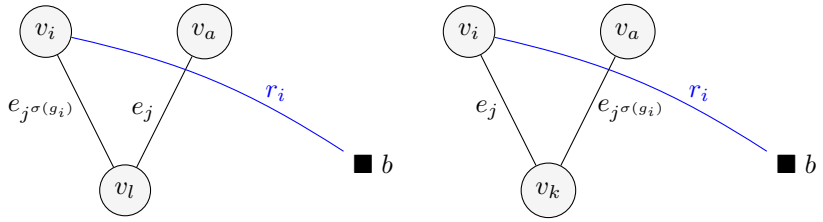
Proof: Suppose $x \in \{k, l\}$ and O is an ordering of the set $\{i, x, a, b\}$. Let O' be the ordering O, except with a and b exchanged. Suppose the generators in the local picture of $g_i|_j$ have order O. Since i, x, a, b are distinct, we may rotate the edge connecting v_a and v_b 180° without influencing the edge connecting v_i and v_x. Now if we swap the labels a and b, the generators of the local picture have order O'. However, since rotation is a continuous deformation, the new loop $g_i|'_j$ must be homotopic to $g_i|_j$. Since we may proceed similarly for every $g_i|_j$ whose associated local picture has generators ordered by either O or O', we

have our result by symmetry. □

This reduces the number of sub-cases to 24, which we classify by using homotopy arguments similar to those described above in Case 1(a), sub-case 3.

Case 2 (b) : $j, j^{\sigma(g_i)} \neq 1$; $v_k = v_b$ or $v_l = v_b$

If $j \neq 1$ is the first number in the monodromy $\sigma(g_i)$, it is possible that the last edge r_i crosses shares an adjacent vertex with the edge $e_{j^{\sigma(g_i)}} \in E$. We let $v_l = v_b$ in this case. Similarly, if $j \neq 1$ is the last number in $\sigma(g_i)$, it is possible that the last edge r_i crosses shares an adjacent vertex with the edge $e_j \in E$. We let $v_k = v_b$ in this case.

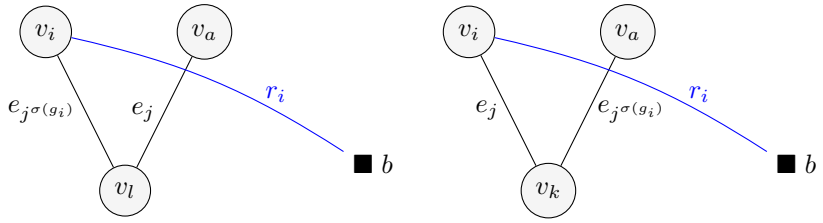


These two cases each have $3! = 6$ sub-cases, arising from the permutations of the sets $\{i, k, a\}$ and $\{i, l, a\}$.

(1) $a < k < i$	(2) $k < a < i$	(3) $a < i < k$	(4) $i < a < k$
(5) $k < i < a$	(6) $i < k < a$	(7) $a < l < i$	(8) $l < a < i$
(9) $a < i < l$	(10) $i < a < l$	(11) $l < i < a$	(12) $i < l < a$

Fortunately, using homotopy arguments similar to those presented above, we find there is only one choice of homotopy class of paths for each generator stem in any one of these twelve sub-cases.

Case 2 (c) : $j = 1$ or $j^{\sigma(g_i)} = 1$



j is first in $\sigma(g_i)$ $\qquad$ j is last in $\sigma(g_i)$

Let $j = 1$. Since neither r_i nor r_l can cross e_j, there is only one choice of homotopy class of paths for each. Since there are $2! = 2$ permutations of the set $\{i, l\}$, there are only two sub-cases of 2(c) for which $j = 1$. By symmetry we have the same for the case $j^{\sigma(g_i)} = 1$ and therefore we conclude that there are exactly four sub-cases of Case 2(c), each with exactly one choice of homotopy class of paths for each generator stem.

Interpreting the Local Picture

Identifying the Local Picture

Identifying the local picture of $g_i|_j$ in our classification is straightforward. If j is neither the first nor last number in the monodromy, then the local picture belongs to Case 1; otherwise it belongs to Case 2. Checking the number of distinct vertices adjacent to e_j and $e_{j^{\sigma(g_i)}}$ further refines this placement to either 1(a), 1(b), 2(a), 2(b), or 2(c). Then the order of the magnitudes of the indices of these adjacent vertices together with the knowledge of which generators in the local picture cross e_j and $e_{j^{\sigma(g_i)}}$ dictates a unique local picture corresponding to $g_i|_j$. A catalogue of these local pictures and characteristics sufficient to find the data (1), (2) and (3) is found in the Appendix. In what follows we calculate (1), (2), and (3) from a local picture $g_i|_j$ and thereby decide what characteristics are necessary to include in the Appendix.

Extracting information from the local picture

From the local picture of $g_i|_j$, we wish to deduce

1. The generators g_s, g_{s+1} that $g_i|_j$ starts "between" and the generators g_e, g_{e+1} that $g_i|_j$ ends "between,"

2. The ordered list $c_1, \ldots, c_k$ of the generators g_i that $g_i|_j$ intersects away from b, written in the order $g_i|_j$ intersects them,

3. The orientation, positive or negative, with which $g_i|_j$ intersects the generator corresponding to c_x, for $1 \leq x \leq k$.

Since we may write $g_i|_j$ as a loop which follows only edges and generators in the local picture of $g_i|_j$, we may let g_α be the generator which $g_i|_j$ follows to the edge e_j and let g_β be the generator $g_i|_j$ follows when leaving e_j. Then the only generators which $g_i|_j$ can cross while traveling along e_j are the generators which cross e_j between the intersection of e_j with g_α and the intersection of e_j with g_β. Since we input for each edge an ordered list of the generators that cross that edge, the generators $g_i|_j$ crosses while following e_j are precisely those in the input for e_j listed between g_α and g_β.

As we can do the same for $e_{j^{\sigma(g_i)}}$, and by Lemma 3 the only generators $g_i|_j$ can cross are those which cross e_j or $e_{j^{\sigma(g_i)}}$, we have found all the generators corresponding to $c_1, \ldots, c_k$. To ensure that these generators are listed in the order $g_i|_j$ intersects them, we make note of the direction $g_i|_j$ travels along e_j and $e_{j^{\sigma(g_i)}}$ and which edge, e_j or $e_{j^{\sigma(g_i)}}$, $g_i|_j$ traverses first, if it does traverse one.

Next we find the orientation with which $g_i|_j$ intersects the generator corresponding to c_x for $1 \leq x \leq k$. First, we consider the c_x's which correspond to generators not present in the local picture of $g_i|_j$. To do this, we let the set of indices of the generators in the local picture partition its complement in the set $\{1, \ldots, n\}$. We refer to this partition as the partition of $\{1, \ldots, n\}$ induced by the local picture of $g_i|_j$, and call the parts of this partition *ranges* T.

Lemma 8: Suppose the indices m, n of two generators g_m, g_n of $\pi_1(S^2 \setminus V, b)$ belong to the same range T in the partition of $\{1, \ldots, n\}$ induced by the local picture of $g_i|_j$. Then if g_m and g_n both intersect e_j or both intersect $e_{j^{\sigma(g_i)}}$, they must do so with the same orientation.

Proof: Let n and m belong to T and suppose for contradiction that g_m and g_n intersect an edge e in the local picture with opposite orientations. Let v and v' be the vertices adjacent to e and let R be a region bounded by $g_v, g_{v'}$, and e. We may assume without loss of generality that g_m is entering R and g_n is exiting R at their respective points of intersection. Since R is bounded by generators and a single edge in in E, g_m cannot begin within R and g_n must begin within R. Since R the generators which bounded R are in the local picture, g_n and g_m cannot begin in the same range T, which is a contradiction.□

Thus for each range T in the partition of $\{1, \ldots, d\}$ induced by the local picture of $g_i|_j$, we need only compute the orientation of a non-empty intersection of a generator in T with e_j and the orientation of a non-empty intersection of a generator in T with $e_{j^{\sigma(g_i)}}$, if such exists. Therefore to determine the orientation associated with c_x, where c_x corresponds to a generator g_i not present in the local picture, we need only check which range i belongs to and which edge g_i crosses in the intersection associated with c_x. By Lemma 8, every generator in this range crossing that edge will do so with the same orientation.

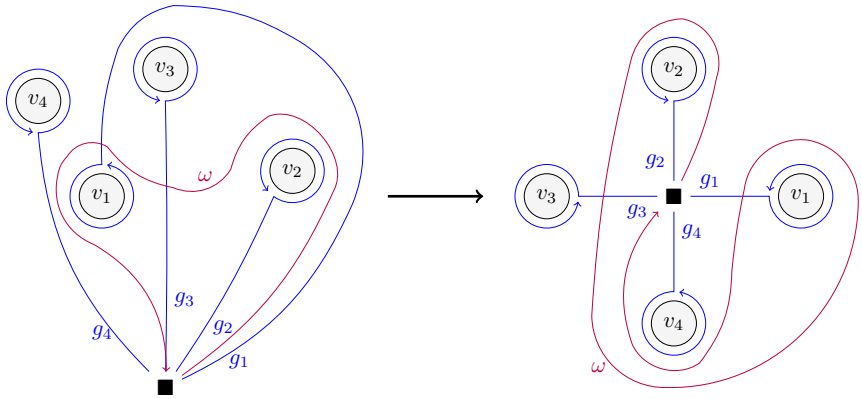
To find the orientation of each c_x corresponding to a generator in the local picture, we check the orientation of the intersection of each generator in the local picture with e_j and $e_{j^{\sigma(g_i)}}$ by hand.

As $\lambda_j \neq \lambda_1$ follows a generator g_α from b counter-clockwise around v_α, we say $g_i|_j$ *starts between* the generators $g_{\alpha-1}$ and g_α since doing so minimizes crossings. Similarly, if $\lambda_{j^{\sigma(g_i)}} \neq \lambda_1$ follows the generator g_β counter-clockwise around v_β, we say $g_i|_j$ *ends between* $g_{\beta-1}$ and g_β. In the case where $j = 1$, $g_i|_j$ follows the path lift $\tilde{g}_{i_{b_j}}$ from b to $b_{j^{\sigma(g_i)}}$ and we say $g_i|_j$ starts between g_{i-1} and g_i. In the case where $j^{\sigma(g_i)} = 1$, $g_i|_j$ follows $\tilde{g}_{i_{b_j}}$ from b_j to b and we say $g_i|_j$ ends between g_i and g_{i+1}.

A catalogue of these local pictures and complete with the results of these computations is found in the Appendix.

Procedure for writing $\omega \in \pi_i(S^2 \setminus V)$ as a word in $g_1, \ldots, g_n$

Suppose ω is a representative of a loop in $\pi_1(S^2 \setminus V, b)$. Since we have assumed the generators g_i of $\pi_1(S^2 \setminus V, b)$ are cyclically ordered and do not cross each other, we may assume, after an appropriate continuous deformation, that the stems r_i of each g_i are straight lines radiating outwards uniformly from the basepoint b.



Suppose that ω leaves the basepoint b between the generators g_s and g_{s+1} and returns to b between the generators g_e and g_{e+1}. Let $c_1, \ldots, c_k$ be the ordered list of generators ω intersects away from b, listed in the order ω intersects them. And suppose we know the orientation with which ω intersects the generator corresponding to c_x for all $1 \leq x \leq k$. Recall that we say ω intersects the generator g_i with *positive orientation* if a person walking along ω would find the basepoint connected to the segment of g_i on the person's left at the point

of intersection. Otherwise, we say ω intersects g_i with *negative orientation*. In the figure above, for example, ω intersects g_3 with positive orientation; on the other hand, it intersects g_1 with negative orientation.

For $1 \leq x \leq k-1$, we divide ω into the $k-1$ paths $w_{x,x+1}$, where $\omega_{x,x+1}$ is the path which follows ω from the point ω intersects the generator corresponding to c_x to the point ω intersects the generator corresponding to c_{x+1}. We let $\omega_{0,1}$ be the path that follows ω from b to the point ω intersects the generator corresponding to c_1, and we let $\omega_{k,k+1}$ be the path that follows ω from the point ω intersects the generator corresponding to c_k to b.

Since ω is homotopic to the concatenation $\omega_{0,1} * \cdots * \omega_{k,k+1}$, if we express each $\omega_{x,x+1}$ for $0 \leq x \leq k$ in terms of the generators $g_1, \ldots, g_n$, we will have an expression in terms of $g_1, \ldots, g_n$ that is homotopic to ω. So we need only find a method for expressing an arbitrary $\omega_{x,x+1}$ in terms of $g_1, \ldots, g_n$.

But this is done easily from the orientations associated with c_x and c_{x+1} alone. In fact, if we let c_x denote the intersection of ω with the generator g_a and c_{x+1} denote the intersection of ω with the generator g_b, then the following table gives an expression for $\omega_{x,x+1}$ in terms of the generators $g_1, \ldots, g_n$ which depends only on the generators ω starts and ends between and the orientations of the intersections associated with c_x and c_{x+1}.

Orientation at c_x	Orientation at c_{x+1}	Solution
$c_x = b$	+1	$g_{s+1} * \cdots * g_{b-1}$
$c_x = b$	-1	$(g_{b+1} * \cdots * g_s)^{-1}$
+1	$c_{x+1} = b$	$g_{a+1} * \cdots * g_e$
-1	$c_{x+1} = b$	$(g_{e+1} * \cdots * g_{a-1})^{-1}$
+1	+1	$g_{a+1} * \cdots * g_{b-1}$
-1	-1	$(g_{b+1} * \cdots * g_{a-1})^{-1}$
+1	-1	$g_{a+1} * \cdots * g_b$
-1	+ 1	$(g_b * \cdots * g_{a-1})^{-1}$

Closing Example

Suppose that we have identified the local picture of a wreath position $g_i|_j$ to be Case 1(a), sub-case $k < i < l$. From the Appendix we retrieve the following information

1. $g_i|_j$ traverses the edge $e_{j^{\sigma(g_i)}}$ from the point at which g_i intersects $e_{j^{\sigma(g_i)}}$ to the point at which g_l intersects $e_{j^{\sigma(g_i)}}$; $g_i|_j$ does not traverse the edge e_j.
2. The orientation of the intersection of a generator g_α with the edge $e_{j^{\sigma(g_i)}}$ is negative in the range $\alpha < k$, negative in the range $k < \alpha < i$, positive in the range $i < \alpha < l$, and again negative in the range $l < \alpha$.

3. $g_i|_j$ intersects none of the generators g_k, g_i, g_l in the local picture.

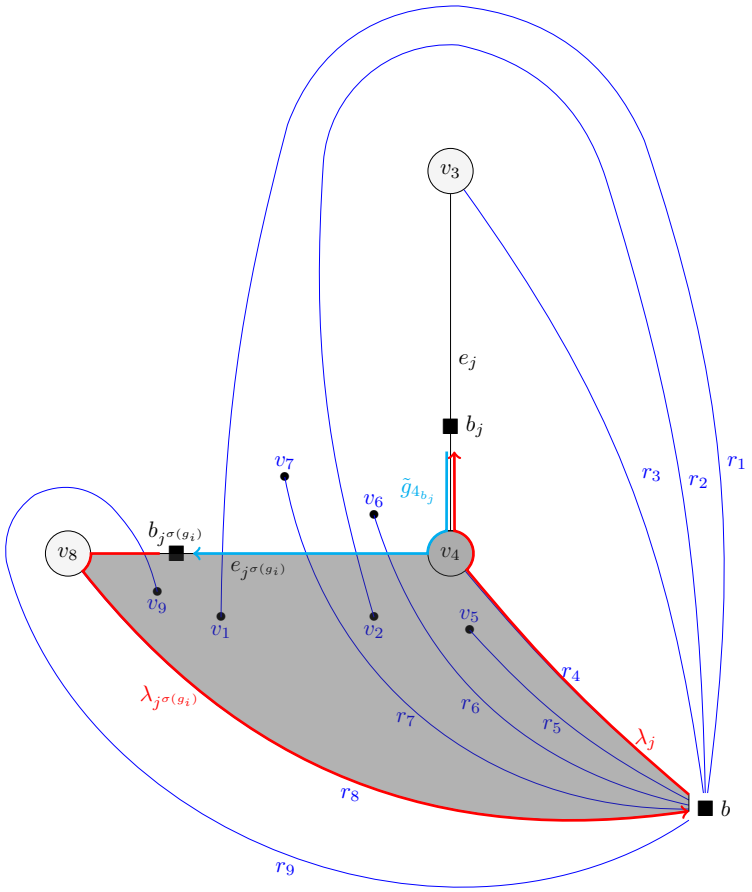
4. $g_i|_j$ starts between the generators g_{i-1} and g_i and ends between the generators g_{l-1} and g_l.

Suppose $k = 3, i = 4$, and $l = 8$ in this example, and suppose $g_4, g_6, g_2, g_7, g_1, g_9, g_8$ is the inputted list of the generators that cross the edge $e_{j^{\sigma(g_i)}}$, ordered from the generator that crosses $e_{j^{\sigma(g_i)}}$ nearest v_4 to the generator that crosses $e_{j^{\sigma(g_i)}}$ nearest v_8. Since $g_i|_j$ traverses $e_{j^{\sigma(g_i)}}$ between the points where g_4 and g_8 cross $e_{j^{\sigma(g_i)}}$ and neither g_4 nor g_8 intersect $g_i|_j$, the list g_6, g_2, g_7, g_1, g_9 is the list of the generators $g_i|_j$ intersects, written in the order $g_i|_j$ intersects them. By examining the ranges their indices belong to, we see that $g_i|_j$ intersects g_2, g_1, and g_9 with negative orientation and g_6 and g_7 with positive orientation. This, along with the fact that $g_i|_j$ starts between g_3 and g_4 and ends between g_7 and g_8, is enough to compute $g_i|_j$ in terms of $g_1, \ldots, g_9$.

Using the table above, we find that

$$g_i|_j = \underbrace{g_4 g_5}_{\omega_{0,1}} \underbrace{g_7 g_8 g_9 g_1 g_2}_{\omega_{1,2}} \underbrace{(g_7 g_8 g_9 g_1)^{-1}}_{\omega_{2,3}} \underbrace{g_8 g_9 g_1}_{\omega_{3,4}} \underbrace{g_8 g_9 g_1 g_2 g_3 g_4 g_5 g_6 g_7 g_8)^{-1}}_{\omega_{5,6}}$$

where $\omega_{4,5}$ is homotopic to the constant loop. We invite the reader to check this both with the process we have described above and by hand with the picture we provide below.



References

[1] David Tischler, *Critical Points and Values of Complex Polynomials*. Journal of Complexity, Volume 5, pp. 438-456 (1989)

[2] Volodymyr Nekrashevych, *Self-similar groups*. Mathematical Surveys and Monographs. Amer. Math. Soc., Volume 117, (2005).

[3] Fu Liu and Brian Osserman, *The irreducibility of certain pure-cycle Hurwitz spaces*. American Journal of Mathematics, Volume 130, Number 6, pp. 1687-1708 (2008)

[4] Kristen Cordwell and Selina Gilbertson, *On the Realizability of Critical Orbit Portraits*. Research Experience for Undergraduates Research Re-

ports Indiana University, Bloomington Summer 2012 pp. 44-60 (2012). http:/www.math.indiana.edu/reu/2012/reu2012.pdf.

[5] Kevin M. Pilgrim and Tan Lei *Combining rational maps and controlling obstructions*. Ergod. Th. & Dynam. Sys., Volume 18, pp. 221-245 (1998).

Guide to the Appendix

The tables in the following sections provide the data necessary to compute the wreath algorithm. Precise definitions for the symbols $i, k, l, a, b, j, n, g_i|_j$ and $j^{\sigma(g_i)}$ can be found in the body of the report.

For each $g_i|_j$, the algorithm first identifies its local picture by its **Subcase** which is an ordering of the indices of the vertices in the local picture: i, k, l for Case 1(a), i, l, a, b for Case 2(a), first in monodromy, i, k, a, b for Case 2(a), last in monodromy, i, l, a for Case 2(b), first in monodromy, i, k, a for Case 2(b), last in monodromy, i, l for Case 2(c), first in monodromy, and i, k for Case 2(c), last in monodromy. These lists correspond to the alphabetical ordering of each Case.

When the **Subcase** is not enough to determine the local picture, the algorithm checks the **Subsubcase**, the final identifier. The **Subsubcase** is determined by which generators in the local picture cross j or $j^{\sigma(g_i)}$. We provide an orientation $\pm$ when a generator crosses $g_i|_j$.

To write $g_i|_j$ in terms of generators $g_1 \dots g_n$, we need to know between which two generators it begins and ends. It begins between g_s and g_{s+1} and ends between g_e and g_{e+1}, all of which are recorded under the column headed **Start/End**.

Next, the algorithm requires the edges $g_i|_j$ traveres and in what order, and the each edge's orientation. Since $g_i|_j$ traverses a maximum of two edges the **Position** is labeled First or Second. When $g_i|_j$ traverses no edges, we indicate that it is not applicable, n/a. These are **Label**ed j or $j^{\sigma(g_i)}$ in the local picture and have orientation $\pm$: $+$ when $g_i|_j$ runs from the lower indexed vertex to the higher indexed vertex and $-$ otherwise. The indices of the **Start** and the **End** vertices also provide the orientation; if **Start**>**End** then the orientation is $-$.

After obtaining an ordered list of generator crosses on $g_i|_j$, the algorithm must determine the orientation of each cross on each edge based on *ranges*. Suppose the **Subcase** was $\alpha < \beta < \gamma$. Then $\{T_1, \dots, T_4\}$ represent the ranges

$$0 < T_1 < \alpha \qquad \alpha < T_2 < \beta \qquad \beta < T_3 < \gamma \qquad \gamma < T_4 < n+1$$

or $0 < T_1 < \alpha < T_2 < \beta < T_3 < \gamma < T_4 < n+1$ and similarly for $\{T_1, \dots T_p\}$ for any p. For each edge in each local picture, the **Orientation** columns provide the orientations for the crosses with ranges. When a generator in a given range cannot cross the given edge, we indicate that it is not possible, n/p. If the cross is a generator in the local picture, the orientation is recorded in the **Subsubcase** column or the **Start** or **End** columns when it is located at the start or end of an edge segment.

Here, as elsewhere, addition is performed modulo n.

A Case 1(a)

j is neither the first not the last number in the monodromy and l and k are distinct

Identify Local Picture		Start/End		Edges Traversed					Orientation			
Subcase	Subsubcase	g_s	g_e	Position	Label	$\pm$	Start	End	T_1	T_2	T_3	T_4
$i<k<l$	g_l crosses j, +	$k-1$	$l-1$	First	j	$-$	k	i	+	$-$	+	+
				Second	$j^{\sigma(g_i)}$	+	i	l	$-$	+	+	$-$
$i<k<l$	g_k crosses $j^{\sigma(g_i)}$, +	$k-1$	$l-1$	First	j	$-$	k	i	+	$-$	+	+
				Second	$j^{\sigma(g_i)}$	+	i	l	$-$	+	+	$-$
$i<l<k$	n/a	$k-1$	$l-1$	First	j	$-$	k	i	+	$-$	$-$	+
				Second	$j^{\sigma(g_i)}$	+	i	l	$-$	+	$-$	$-$
$k<i<l$	n/a	$i-1$	$l-1$	First	$j^{\sigma(g_i)}$	+	i	l	$-$	$-$	+	$-$
$k<l<i$	g_l crosses j	$i-1$	$i-1$	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
$k<l<i$	g_k crosses $j^{\sigma(g_i)}$	$i-1$	$i-1$	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
$l<i<k$	g_l crosses j, +	$k-1$	$i-1$	First	j	$-$	k	i	+	+	$-$	+
$l<i<k$	g_k crosses $j^{\sigma(g_i)}$	$k-1$	$i-1$	First	j	$-$	k	i	+	+	$-$	+
$l<k<i$	n/a	$i-1$	$i-1$	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

B Case 2(b)

j is neither the first not the last in the monodromy and $k=l$

Case 2(b) always gives the constant loop, denoted e.

C Case 2(a)

First in monodromy

j is the first number in the monodromy, $j \neq 1$ and l and b are distinct.

Identify Local Picture		Start/End		Edges Traversed					Orientation				
Subcase	Subsubcase	g_s	g_e	Position	Label	$\pm$	Start	End	T_1	T_2	T_3	T_4	T_5
$i<l<a<b$	g_l crosses j	$b-1$	$l-1$	First	j	$-$	b	i	$+$	n/p	n/p	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$+$	$-$	$-$	$-$
$i<l<a<b$	g_a crosses $j^{\sigma(g_i)}$, $-$	$b-1$	$l-1$	First	j	$-$	b	i	$+$	n/p	n/p	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$+$	$-$	$-$	$-$
$i<l<b<a$	g_l crosses j	$a-1$	$l-1$	First	j	$-$	a	i	$+$	n/p	n/p	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$+$	$-$	$-$	$-$
$i<l<b<a$	g_b crosses $j^{\sigma(g_i)}$, $-$	$a-1$	$l-1$	First	j	$-$	a	i	$+$	n/p	n/p	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$+$	$-$	$-$	$-$
$i<a<l<b$	n/a	$b-1$	$l-1$	First	j	$-$	b	i	$+$	n/p	$-$	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$+$	$+$	$-$	$-$
$i<a<b<l$	g_l crosses j, $+$	$b-1$	$l-1$	First	j	$-$	b	i	$+$	n/p	$-$	$+$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$+$	$+$	$+$	$-$
$i<a<b<l$	g_b crosses $j^{\sigma(g_i)}$	$b-1$	$l-1$	First	j	$-$	b	i	$+$	n/p	$-$	$+$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$+$	$+$	$+$	$-$
$i<b<l<a$	n/a	$a-1$	$l-1$	First	j	$-$	a	i	$+$	n/p	$-$	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$+$	$+$	$-$	$-$
$i<b<a<l$	g_l crosses j, $+$	$a-1$	$l-1$	First	j	$-$	a	i	$+$	n/p	$-$	$+$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$+$	$+$	$+$	$-$
$i<b<a<l$	g_a crosses $j^{\sigma(g_i)}$	$a-1$	$l-1$	First	j	$-$	a	i	$+$	n/p	$-$	$+$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$+$	$+$	$+$	$-$
$l<i<a<b$	g_l crosses j, $+$	$b-1$	$i-1$	First	j	$-$	b	i	$+$	$+$	n/p	$-$	$+$
$l<i<a<b$	g_b crosses $j^{\sigma(g_i)}$	$b-1$	$i-1$	First	j	$-$	b	i	$+$	$+$	n/p	$-$	$+$
$l<i<b<a$	g_l crosses j, $+$	$a-1$	$i-1$	First	j	$-$	a	i	$+$	$+$	n/p	$-$	$+$
$l<i<b<a$	g_a crosses $j^{\sigma(g_i)}$	$a-1$	$i-1$	First	j	$-$	a	i	$+$	$+$	n/p	$-$	$+$
$l<a<i<b$	n/a	$b-1$	$i-1$	First	j	$-$	b	$i-$	$+$	$+$	n/p	$-$	$+$
$l<a<b<i$	g_l crosses j	$b-1$	$i-1$	First	j	$-$	b	i	n/p	n/p	$-$	$+$	n/p
$l<a<b<i$	g_a crosses $j^{\sigma(g_i)}$	$b-1$	$i-1$	First	j	$-$	b	i	n/p	n/p	$-$	$+$	n/p
$l<b<i<a$	n/a	$a-1$	$i-1$	First	j	$-$	a	$i-$	$+$	$+$	n/p	$-$	$+$
$l<b<a<i$	g_l crosses j	$a-1$	$i-1$	First	j	$-$	a	i	n/p	n/p	$-$	$+$	n/p
$l<b<a<i$	g_b crosses $j^{\sigma(g_i)}$	$a-1$	$i-1$	First	j	$-$	a	i	n/p	n/p	$-$	$+$	n/p

Identify Local Picture		Start/End		Edges Traversed					Orientation				
Subcase	Subsubcase	g_s	g_e	Position	Label	$\pm$	Start	End	T_1	T_2	T_3	T_4	T_5
$a<i<l<b$	g_l crosses j, $-$	$b-1$	$l-1$	First	j	$-$	b	$i-$	$+$	n/p	$-$	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$-$	$+$	$-$	$-$
$a<i<l<b$	g_b crosses $j^{\sigma(g_i)}$, $-$	$b-1$	$l-1$	First	j	$-$	b	$i-$	$+$	n/p	$-$	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$-$	$+$	$-$	$-$
$a<i<b<l$	n/a	$b-1$	$l-1$	First	j	$-$	b	$i-$	$+$	n/p	$-$	$+$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$-$	$+$	$+$	$-$
$a<l<i<b$	g_l crosses j	$b-1$	$i-1$	First	j	$-$	b	i	$+$	n/p	n/p	$-$	$+$
$a<l<i<b$	g_a crosses $j^{\sigma(g_i)}$	$b-1$	$i-1$	First	j	$-$	b	i	$+$	n/p	n/p	$-$	$+$
$a<l<b<i$	n/a	$b-1$	$i-1$	First	j	$-$	b	i	n/p	$-$	$-$	$+$	n/p
$a<b<i<l$	g_l crosses j	$b-1$	$l-1$	First	j	$-$	b	i	n/p	$-$	$+$	n/p	n/p
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$-$	$-$	$+$	$-$
$a<b<i<l$	g_a crosses $j^{\sigma(g_i)}$, $-$	$b-1$	$l-1$	First	j	$-$	b	i	n/p	$-$	$+$	n/p	n/p
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$-$	$-$	$+$	$-$
$a<b<l<i$	g_a crosses j, $+$	$b-1$	$i-1$	First	j	$-$	b	i	n/p	$-$	$+$	$+$	n/p
$a<b<l<i$	g_b crosses $j^{\sigma(g_i)}$	$b-1$	$i-1$	First	j	$-$	b	i	n/p	$-$	$+$	$+$	n/p
$b<i<l<a$	g_l crosses j, $-$	$a-1$	$l-1$	First	j	$-$	a	$i-$	$+$	n/p	$-$	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$-$	$+$	$-$	$-$
$b<i<l<a$	g_a crosses $j^{\sigma(g_i)}$, $-$	$a-1$	$l-1$	First	j	$-$	a	$i-$	$+$	n/p	$-$	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$-$	$+$	$-$	$-$
$b<i<a<l$	n/a	$a-1$	$l-1$	First	j	$-$	a	$i-$	$+$	n/p	$-$	$+$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$-$	$+$	$+$	$-$
$b<l<i<a$	g_l crosses j	$a-1$	$i-1$	First	j	$-$	a	i	$+$	n/p	n/p	$-$	$+$
$b<l<i<a$	g_b crosses $j^{\sigma(g_i)}$	$a-1$	$i-1$	First	j	$-$	a	i	$+$	n/p	n/p	$-$	$+$
$b<l<a<i$	n/a	$a-1$	$i-1$	First	j	$-$	a	i	n/p	$-$	$-$	$+$	n/p
$b<a<i<l$	g_l crosses j	$a-1$	$l-1$	First	j	$-$	a	i	n/p	$-$	$+$	n/p	n/p
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$-$	$-$	$+$	$-$
$b<a<i<l$	g_b crosses $j^{\sigma(g_i)}$, $-$	$a-1$	$l-1$	First	j	$-$	a	i	n/p	$-$	$+$	n/p	n/p
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$-$	$-$	$+$	$-$
$b<a<l<i$	g_b crosses j, $+$	$a-1$	$i-1$	First	j	$-$	a	i	n/p	$-$	$+$	$+$	n/p
$b<a<l<i$	g_a crosses $j^{\sigma(g_i)}$	$a-1$	$i-1$	First	j	$-$	a	i	n/p	$-$	$+$	$+$	n/p

Last in monodromy

j is the last number in the monodromy, $j^{\sigma(g_i)} \neq 1$ and k and b are distinct.

Identify Local Picture		Start/End		Edges Traversed					Orientation				
Subcase	Subsubcase	g_s	g_e	Position	Label	$\pm$	Start	End	T_1	T_2	T_3	T_4	T_5
$i<k<a<b$	g_k crosses j	$k-1$	$b-1$	First	j	$-$	l	i	$+$	$-$	$+$	$+$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	$i-$	b	$-$	n/p	n/p	$+$	$-$
$i<k<a<b$	g_a crosses $j^{\sigma(g_i)}$, $+$	$k-1$	$b-1$	First	j	$-$	l	i	$+$	$-$	$+$	$+$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	$i-$	b	$-$	n/p	n/p	$+$	$-$
$i<k<b<a$	g_k crosses j	$k-1$	$a-1$	First	j	$-$	k	i	$+$	$-$	$+$	$+$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	$i-$	a	$-$	n/p	n/p	$+$	$-$
$i<k<b<a$	g_b crosses $j^{\sigma(g_i)}$, $+$	$k-1$	$a-1$	First	j	$-$	k	i	$+$	$-$	$+$	$+$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	$i-$	a	$-$	n/p	n/p	$+$	$-$
$i<a<k<b$	n/a	$k-1$	$b-1$	First	j	$-$	k	i	$+$	$-$	$-$	$+$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	$i-$	b	$-$	n/p	$+$	$+$	$-$
$i<a<b<k$	g_k crosses j, $-$	$k-1$	$b-1$	First	j	$-$	k	i	$+$	$-$	$-$	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	$i-$	b	$-$	n/p	$+$	$-$	$-$
$i<a<b<k$	g_b crosses $j^{\sigma(g_i)}$	$k-1$	$b-1$	First	j	$-$	l	i	$+$	$-$	$-$	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	$i-$	b	$-$	n/p	$+$	$-$	$-$
$i<b<k<a$	n/a	$k-1$	$a-1$	First	j	$-$	k	i	$+$	$-$	$-$	$+$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	$i-$	a	$-$	n/p	$+$	$+$	$-$
$i<b<a<k$	g_l crosses j, $-$	$k-1$	$a-1$	First	j	$-$	k	i	$+$	$-$	$-$	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	$i-$	a	$-$	n/p	$+$	$-$	$-$
$i<b<a<k$	g_a crosses $j^{\sigma(g_i)}$	$k-1$	$a-1$	First	j	$-$	k	i	$+$	$-$	$-$	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	$i-$	a	$-$	n/p	$+$	$-$	$-$
$k<i<a<b$	g_k crosses j, $-$	$i-1$	$b-1$	First	$j^{\sigma(g_i)}$	$+$	$i-$	b	$-$	$-$	n/p	$+$	$-$
$k<i<a<b$	g_b crosses $j^{\sigma(g_i)}$	$i-1$	$b-1$	First	$j^{\sigma(g_i)}$	$+$	$i-$	b	$-$	$-$	n/p	$+$	$-$
$k<i<b<a$	g_k crosses j, $-$	$i-1$	$a-1$	First	$j^{\sigma(g_i)}$	$+$	$i-$	a	$-$	$-$	n/p	$+$	$-$
$k<i<b<a$	g_a crosses $j^{\sigma(g_i)}$	$i-1$	$a-1$	First	$j^{\sigma(g_i)}$	$+$	$i-$	a	$-$	$-$	n/p	$+$	$-$
$k<a<i<b$	n/a	$i-1$	$b-1$	First	$j^{\sigma(g_i)}$	$+$	i	b	$-$	$-$	n/p	$+$	$-$
$k<a<b<i$	g_k crosses j	$i-1$	$b-1$	First	$j^{\sigma(g_i)}$	$+$	$i-$	b	n/p	n/p	$+$	$-$	n/p
$k<a<b<i$	g_a crosses $j^{\sigma(g_i)}$	$i-1$	$b-1$	First	$j^{\sigma(g_i)}$	$+$	$i-$	b	n/p	n/p	$+$	$-$	n/p
$k<b<i<a$	n/a	$i-1$	$a-1$	First	$j^{\sigma(g_i)}$	$+$	i	a	$-$	$-$	n/p	$+$	$-$
$k<b<a<i$	g_k crosses j	$i-1$	$a-1$	First	$j^{\sigma(g_i)}$	$+$	$i-$	a	n/p	n/p	$+$	$-$	n/p
$k<b<a<i$	g_b crosses $j^{\sigma(g_i)}$	$i-1$	$a-1$	First	$j^{\sigma(g_i)}$	$+$	$i-$	a	n/p	n/p	$+$	$-$	n/p

Identify Local Picture		Start/End		Edges Traversed					Orientation				
Subcase	Subsubcase	g_s	g_e	Position	Label	$\pm$	Start	End	T_1	T_2	T_3	T_4	T_5
$a<i<k<b$	g_k crosses j, +	$k-1$	$b-1$	First	j	$-$	k	i	+	+	$-$	+	+
				Second	$j^{\sigma(g_i)}$	+	i	b	$-$	n/p	+	+	$-$
$a<i<k<b$	g_b crosses $j^{\sigma(g_i)}$, +	$k-1$	$b-1$	First	j	$-$	k	i	+	+	$-$	+	+
				Second	$j^{\sigma(g_i)}$	+	i	b	$-$	n/p	+	+	$-$
$a<i<b<k$	n/a	$k-1$	$b-1$	First	j	$-$	k	i	+	+	$-$	$-$	+
				Second	$j^{\sigma(g_i)}$	+	i	b	$-$	n/p	+	$-$	$-$
$a<k<i<b$	g_k crosses j	$i-1$	$b-1$	First	$j^{\sigma(g_i)}$	+	i	b	$-$	n/p	n/p	+	$-$
$a<k<i<b$	g_a crosses $j^{\sigma(g_i)}$	$i-1$	$b-1$	First	$j^{\sigma(g_i)}$	+	i	b	$-$	n/p	n/p	+	$-$
$a<k<b<i$	n/a	$i-1$	$b-1$	First	$j^{\sigma(g_i)}$	+	$i-$	b	n/p	+	+	$-$	n/p
$a<b<i<k$	g_k crosses j	$k-1$	$b-1$	First	j	$-$	k	i	+	+	+	$-$	+
				Second	$j^{\sigma(g_i)}$	+	$i-$	b	n/p	+	$-$	n/p	n/p
$a<b<i<k$	g_a crosses $j^{\sigma(g_i)}$, +	$k-1$	$b-1$	First	j	$-$	k	i	+	+	+	$-$	+
				Second	$j^{\sigma(g_i)}$	+	$i-$	b	n/p	+	$-$	n/p	n/p
$a<b<k<i$	g_a crosses j, $-$	$i-1$	$b-1$	First	$j^{\sigma(g_i)}$	+	$i-$	a	n/p	+	$-$	$-$	n/p
$a<b<k<i$	g_b crosses $j^{\sigma(g_i)}$	$i-1$	$b-1$	First	$j^{\sigma(g_i)}$	+	$i-$	b	n/p	+	$-$	$-$	n/p
$b<i<k<a$	g_k crosses j, +	$k-1$	$a-1$	First	j	$-$	k	i	+	+	$-$	+	+
				Second	$j^{\sigma(g_i)}$	+	i	a	$-$	n/p	+	+	$-$
$b<i<k<a$	g_a crosses $j^{\sigma(g_i)}$, +	$k-1$	$a-1$	First	j	$-$	k	i	+	+	$-$	+	+
				Second	$j^{\sigma(g_i)}$	+	i	a	$-$	n/p	+	+	$-$
$b<i<a<k$	n/a	$k-1$	$a-1$	First	j	$-$	k	i	+	+	$-$	$-$	+
				Second	$j^{\sigma(g_i)}$	+	i	a	$-$	n/p	+	$-$	$-$
$b<k<i<a$	g_k crosses j	$i-1$	$a-1$	First	$j^{\sigma(g_i)}$	+	i	a	$-$	n/p	n/p	+	$-$
$b<k<i<a$	g_b crosses $j^{\sigma(g_i)}$	$i-1$	$a-1$	First	$j^{\sigma(g_i)}$	+	i	a	$-$	n/p	n/p	+	$-$
$b<k<a<i$	n/a	$i-1$	$a-1$	First	$j^{\sigma(g_i)}$	+	$i-$	a	n/p	+	+	$-$	n/p
$b<a<i<k$	g_l crosses j	$k-1$	$a-1$	First	j	$-$	k	i	+	+	+	$-$	+
				Second	$j^{\sigma(g_i)}$	+	$i-$	a	n/p	+	$-$	n/p	n/p
$b<a<i<k$	g_b crosses $j^{\sigma(g_i)}$, +	$k-1$	$a-1$	First	j	$-$	k	i	+	+	+	$-$	+
				Second	$j^{\sigma(g_i)}$	+	$i-$	a	n/p	+	$-$	n/p	n/p
$b<a<k<i$	g_b crosses j, $-$	$i-1$	$a-1$	First	$j^{\sigma(g_i)}$	+	$i-$	a	n/p	+	$-$	$-$	n/p
$b<a<k<i$	g_a crosses $j^{\sigma(g_i)}$	$i-1$	$a-1$	First	$j^{\sigma(g_i)}$	+	$i-$	a	n/p	+	$-$	$-$	n/p

D Case 2(b)

First in monodromy

j is the first number in the monodromy, $j \neq 1$, $b = l$ and it is denoted l

Identify Local Picture		Start/End		Edges Traversed					Orientation			
Subcase	Subsubcase	g_s	g_e	Position	Label	$\pm$	Start	End	T_1	T_2	T_3	T_4
$i < l < a$	n/a	$a-1$	$l-1$	First	j	$-$	a	i	$+$	n/p	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$+$	$-$	$-$
$i < a < l$	n/a	$l-1$	$l-1$	First	j	$-$	l	i	$+$	n/p	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$+$	$+$	$-$
$l < i < a$	n/a	$a-1$	$i-1$	First	j	$-$	a	$i-$	$+$	n/p	$-$	$+$
$l < a < i$	n/a	$a-1$	$i-1$	First	j	$+$	a	i	$-$	$-$	$+$	$-$
$a < i < l$	n/a	$k-1$	$l-1$	First	j	$-$	l	$i-$	$+$	n/p	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$-$	$+$	$-$
$a < l < i$	n/a	$l-1$	$i-1$	First	j	$-$	l	i	n/p	$-$	$+$	n/p

Last in monodromy

j is the last number in the monodromy, $j^{\sigma(g_i)} \neq 1$, $b = k$ and it is denoted k

Identify Local Picture		Start/End		Edges Traversed					Orientation			
Subcase	Subsubcase	g_s	g_e	Position	Label	$\pm$	Start	End	T_1	T_2	T_3	T_4
$i < k < a$	n/a	$k-1$	$a-1$	First	j	$-$	k	i	$+$	$-$	$+$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	$i-$	a	$-$	n/p	$+$	$-$
$i < a < k$	n/a	$k-1$	$k-1$	First	j	$-$	k	i	$+$	$-$	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	$i-$	k	n/p	$+$	$-$	n/p
$k < i < a$	n/a	$i-1$	$a-1$	First	$j^{\sigma(g_i)}$	$+$	i	a	$-$	$-$	$+$	$-$
$k < a < i$	n/a	$i-1$	$a-1$	First	$j^{\sigma(g_i)}$	$-$	$i-$	a	$+$	$+$	$-$	$+$
$a < i < k$	n/a	$k-1$	$k-1$	First	j	$-$	k	i	$+$	$+$	$-$	$+$
				Second	$j^{\sigma(g_i)}$	$+$	i	k	n/p	$+$	$-$	n/p
$a < k < i$	n/a	$i-1$	$k-1$	First	$j^{\sigma(g_i)}$	$-$	$i-$	k	n/p	$+$	$-$	n/p

E Case 2(c)

First in monodromy

j is the first number in the monodromy and $j = 1$

Identify Local Picture		Start/End		Edges Traversed					Orientation		
Subcase	Subsubcase	g_s	g_e	Position	Label	$\pm$	Start	End	r_1	r_2	r_3
$i < l$	n/a	$i-1$	$l-1$	First	$j^{\sigma(g_i)}$	$+$	i	l	$-$	$+$	$-$
$l < i$	n/a	$i-1$	$i-1$	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Last in monodromy

j is the last number in the monodromy and $j^{\sigma(g_i)} = 1$

Identify Local Picture		Start/End		Edges Traversed					Orientation		
Subcase	Subsubcase	g_s	g_e	Position	Label	$\pm$	Start	End	r_1	r_2	r_3
$i < k$	n/a	$k-1$	i	First	j	$-$	k	i	$+$	$-$	$+$
$k < i$	n/a	$i-1$	i	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a