

involve

a journal of mathematics

Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

Colin Adams	David Larson
John V. Baxley	Suzanne Lenhart
Arthur T. Benjamin	Chi-Kwong Li
Martin Bohner	Robert B. Lund
Nigel Boston	Gaven J. Martin
Amarjit S. Budhiraja	Mary Meyer
Pietro Cerone	Emil Minchev
Scott Chapman	Frank Morgan
Jem N. Corcoran	Mohammad Sal Moslehian
Toka Diagana	Zuhair Nashed
Michael Dorff	Ken Ono
Sever S. Dragomir	Timothy E. O'Brien
Behrouz Emamizadeh	Joseph O'Rourke
Joel Foisy	Yuval Peres
Errin W. Fulp	Y.-F. S. Pétermann
Joseph Gallian	Robert J. Plemmons
Stephan R. Garcia	Carl B. Pomerance
Anant Godbole	Bjorn Poonen
Ron Gould	József H. Przytycki
Andrew Granville	Richard Rebarber
Jerrold Griggs	Robert W. Robinson
Sat Gupta	Filip Saidak
Jim Haglund	James A. Sellers
Johnny Henderson	Andrew J. Sterge
Jim Hoste	Ann Trenk
Natalia Hritonenko	Ravi Vakil
Glenn H. Hurlbert	Antonia Vecchio
Charles R. Johnson	Ram U. Verma
K. B. Kulasekera	John C. Wierman
Gerry Ladas	Michael E. Zieve



involve

msp.org/involve

INVOLVE YOUR STUDENTS IN RESEARCH

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

MANAGING EDITOR

Kenneth S. Berenhaut Wake Forest University, USA

BOARD OF EDITORS

Colin Adams	Williams College, USA	Suzanne Lenhart	University of Tennessee, USA
John V. Baxley	Wake Forest University, NC, USA	Chi-Kwong Li	College of William and Mary, USA
Arthur T. Benjamin	Harvey Mudd College, USA	Robert B. Lund	Clemson University, USA
Martin Bohner	Missouri U of Science and Technology, USA	Gaven J. Martin	Massey University, New Zealand
Nigel Boston	University of Wisconsin, USA	Mary Meyer	Colorado State University, USA
Amarjit S. Budhiraja	U of North Carolina, Chapel Hill, USA	Emil Minchev	Ruse, Bulgaria
Pietro Cerone	La Trobe University, Australia	Frank Morgan	Williams College, USA
Scott Chapman	Sam Houston State University, USA	Mohammad Sal Moslehian	Ferdowsi University of Mashhad, Iran
Joshua N. Cooper	University of South Carolina, USA	Zuhair Nashed	University of Central Florida, USA
Jem N. Corcoran	University of Colorado, USA	Ken Ono	Emory University, USA
Toka Diagana	Howard University, USA	Timothy E. O'Brien	Loyola University Chicago, USA
Michael Dorff	Brigham Young University, USA	Joseph O'Rourke	Smith College, USA
Sever S. Dragomir	Victoria University, Australia	Yuval Peres	Microsoft Research, USA
Behrouz Emamizadeh	The Petroleum Institute, UAE	Y.-F. S. Pétermann	Université de Genève, Switzerland
Joel Foisy	SUNY Potsdam, USA	Robert J. Plemmons	Wake Forest University, USA
Errin W. Fulp	Wake Forest University, USA	Carl B. Pomerance	Dartmouth College, USA
Joseph Gallian	University of Minnesota Duluth, USA	Vadim Ponomarenko	San Diego State University, USA
Stephan R. Garcia	Pomona College, USA	Bjorn Poonen	UC Berkeley, USA
Anant Godbole	East Tennessee State University, USA	James Propp	U Mass Lowell, USA
Ron Gould	Emory University, USA	József H. Przytycki	George Washington University, USA
Andrew Granville	Université Montréal, Canada	Richard Rebarber	University of Nebraska, USA
Jerold Griggs	University of South Carolina, USA	Robert W. Robinson	University of Georgia, USA
Sat Gupta	U of North Carolina, Greensboro, USA	Filip Saidak	U of North Carolina, Greensboro, USA
Jim Haglund	University of Pennsylvania, USA	James A. Sellers	Penn State University, USA
Johnny Henderson	Baylor University, USA	Andrew J. Sterge	Honorary Editor
Jim Hoste	Pitzer College, USA	Ann Trenk	Wellesley College, USA
Natalia Hritonenko	Prairie View A&M University, USA	Ravi Vakil	Stanford University, USA
Glenn H. Hurlbert	Arizona State University, USA	Antonia Vecchio	Consiglio Nazionale delle Ricerche, Italy
Charles R. Johnson	College of William and Mary, USA	Ram U. Verma	University of Toledo, USA
K. B. Kulasekera	Clemson University, USA	John C. Wierman	Johns Hopkins University, USA
Gerry Ladas	University of Rhode Island, USA	Michael E. Zieve	University of Michigan, USA

PRODUCTION

Silvio Levy, Scientific Editor


Cover: Alex Scorpan

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2017 is US \$175/year for the electronic version, and \$235/year (+\$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

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

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

PUBLISHED BY

 **mathematical sciences publishers**
nonprofit scientific publishing

<http://msp.org/>

© 2017 Mathematical Sciences Publishers

Stability analysis for numerical methods applied to an inner ear model

Kimberley Lindenberg, Kees Vuik and Pieter W. J. van Hengel

(Communicated by Kenneth S. Berenhaut)

Diependaal, Duifhuis, Hoogstraten and Viergever investigated three time-integration methods to solve a simplified one-dimensional model of the human cochlea. Two of these time-integration methods are dealt with in this paper, namely fourth-order Runge–Kutta and modified Sielecki. The stability of these two methods is examined, both theoretically and experimentally. This leads to the conclusion that in the case of the fourth-order Runge–Kutta method, a bigger time step can be used in comparison to the modified Sielecki method. This corresponds with the conclusion drawn in the article by Diependaal, Duifhuis, Hoogstraten and Viergever.

1. Introduction

1.1. Motivation. Deafness can be caused by a problem with the mechanical part of the human ear, which consists of three parts, namely the outer ear, the middle ear and the inner ear. The inner ear includes the cochlea (the organ of hearing) and the vestibular system (balance). The cochlea converts incoming sounds into electrochemical (nerve) impulses.

Almost always, it is possible to improve the hearing of those who are hearing impaired. Therefore it is important that deafness or hearing impairment is detected as early as possible. To test the functioning of the hearing system, subjective thresholds are determined at standardised frequencies and are related to standardised average thresholds. However, in general these tests cannot be performed on everyone. For example, they cannot be administered to people who are incapable of responding, such as babies and young children. Besides that, this method tests the functioning of the entire hearing system, not only of the cochlea. This leads to a different problem, because to improve the diagnosis of a hearing deficit it would be useful to separate the functioning of the cochlea from the neural processing [van Hengel 1996].

MSC2010: primary 65L06, 65L07, 65L20, 65M12; secondary 65L05, 65L10.

Keywords: inner ear model, numerical time-integration method, modified Sielecki, fourth-order Runge–Kutta, numerically stable, numerically unstable .

There is an objective test (diagnosing cochlear dysfunction) which has the potential to detect deafness or hearing impairment as early as possible. In this test otoacoustic emissions play an important role. Otoacoustic emissions are very weak sounds produced by the cochlea, in response to stimulation or spontaneously. These sounds can be measured with a sensitive microphone in the ear canal. These otoacoustic emissions give directly measurable information about the condition of the cochlea, and thus can be used when diagnosing cochlear dysfunction. It is known that subjects with cochlear hearing deficits have emissions that differ from those found in people with normal hearing. Since otoacoustic emissions can be directly linked to cochlear functioning, it is possible for objective tests to be carried out on anyone, including babies and small children. The problem is that it remains difficult to link otoacoustic emission levels to cochlear functioning. A deeper understanding of the generative mechanism(s) is thus required. Since the cochlea is extremely vulnerable and difficult to access, *in vivo* studies on otoacoustic emissions cannot be performed in humans. However, these studies are performed in animals to help understand the phenomenon. Additionally, cochlea models are used to study otoacoustic emissions [van Hengel 1996].

1.2. *Early work.* The model used in this paper is obtained from an internal report by Marc van den Raadt, in which the numerical treatment of motion equations is described in detail and which is partly based on the paper by Diependaal et al. [1987], where they examined three time-integration methods (Heun, fourth-order Runge–Kutta and modified Sielecki) in order to solve their model. They also dealt with the numerical stability of these three methods. The time-integration methods have to be numerically stable and this limits the size of the time step used for a given problem.

1.3. *What is new in this paper?* There exist two kinds of stability, analytical and numerical. It is possible that a second-order differential equation is analytically stable (positive damping), but at the same time the used numerical method can be unstable, because too large a step size is used or an improper time-integration method is applied. While most authors examine only the analytical stability, we consider the numerical stability as well and realize that these two kinds of stability are not the same. Diependaal et al. seem to make this distinction between analytical and numerical stability as well. However, their stability analysis is limited to a numerical test (determining the bounds in an experimental way), and they obtain a conservative guess of the step-size limit for each time-integration method by testing different step sizes. Using these numerical tests, Diependaal et al. conclude that in the fourth-order Runge–Kutta method, a bigger time step can be used in comparison with Heun and modified Sielecki methods. In this paper we only examine two time-integration methods, fourth order Runge–Kutta and modified Sielecki, and a real numerical stability analysis is conducted. So, the theoretical bounds for the

time steps are derived and verified with a numerical test. As far as the authors know, the method used to investigate the stability of the modified Sielecki method is not known in the literature, and therefore a new contribution of this paper.

The goal is also to derive a method for stability analysis on the model with parameter variations to simulate hearing loss.

1.4. Structure of the paper. In Section 2, the biological background, the mathematical model and the discretisation of this model are dealt with. The numerical methods used during the project and their properties are examined in Section 3. The numerical experiments are examined in Section 4. Finally, in Section 5 various conclusions are drawn.

2. Problem definition

2.1. Biological background. The cochlea plays an important part in the processing of incoming sounds. The incoming sound waves behave like pressure waves in the ear. The pressure waves, which reach the eardrum, are transmitted via vibrations of the middle ear ossicles to the oval window at the base of the cochlea. These vibrations move the cochlear fluids, which stimulate tiny hair cells on the cochlear partition. Individual hair cells respond to specific sound frequencies so that, depending on the frequency of the sound, only certain hair cells are stimulated [Robles and Ruggero 2001; Bell 2004].

The cochlea looks like a coiled tube. Mechanically this tube is divided into two compartments by the cochlear partition, consisting of the basilar membrane and the organ of Corti (the unfolded cochlea is shown in Figure 1). The two compartments are filled with fluid, which will have equivalent mechanical properties to water in this proposed model. The organ of Corti, which is a cellular layer on the basilar membrane, contains the hair cells that start to move when sound waves enter the

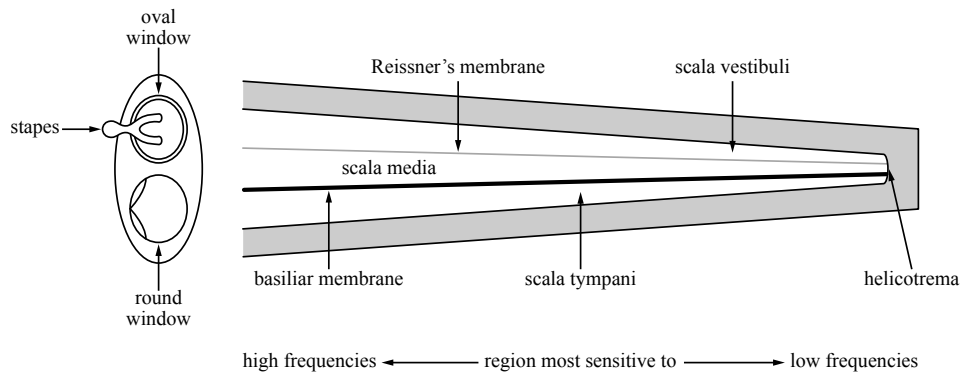


Figure 1. The unfolded cochlea.

ear. The part of the cochlear partition situated at the base will resonate at higher frequencies and the part at the apex will resonate at lower frequencies [Robles and Ruggero 2001; Bell 2004].

2.2. Mathematical model. In our model, the cochlea will be viewed as a straight cylinder. The cochlea will be divided by the cochlear partition into two fluid channels of the same height and the same width. We will also assume that the basilar membrane has the same width from the base to the apex. It is assumed that the base is located at the left side of the cylinder and the apex at the right side.

The one-dimensional cochlea model [Diependaal et al. 1987] is defined by

$$\frac{\partial^2 p}{\partial x^2}(x, t) - \frac{2\rho}{h} \frac{\partial^2 y}{\partial t^2}(x, t) = 0, \quad 0 \leq x \leq L, \quad t \geq 0, \quad (1)$$

where the transmembrane fluid pressure $p(x, t)$ can be written as

$$p(x, t) = m_s \ddot{y}(x, t) + d_s(x) \dot{y}(x, t) + s_s(x) y(x, t).$$

For simplicity, this pressure is assumed to be equal to zero at the helicotrema (at $x = L$). Note that this assumption is an approximation because in reality there can be a small fluid flow as a result of the remaining pressure difference and some damping that affects the flow. For high frequencies, this is almost negligible, but it can play an important role for frequencies below 1 kHz. Equation (1) describes the movement of a cochlear section, and in this equation ρ stands for the density of the cochlear fluid and h is the height of a scala. Here $y(x, t)$ is the excitation in the oscillators, m_s the specific acoustic mass of the basilar membrane, d_s the specific acoustic damping of the basilar membrane and s_s the specific acoustic stiffness of the basilar membrane. Both d_s and s_s vary with the placement of an oscillator.

2.3. Spatial discretisation. If we define $G(x, t) = d_s(x) \dot{y}(x, t) + s_s(x) y(x, t)$, then $p(x, t) - G(x, t) = m_s \ddot{y}(x, t)$. The differential equation (1) can be written as

$$-\frac{\partial^2 p}{\partial x^2}(x, t) + \frac{2\rho}{hm} p(x, t) = \frac{2\rho}{hm} G(x, t), \quad 0 \leq x \leq L, \quad t \geq 0. \quad (2)$$

This model is used to describe $N + 1$ individual cochlear sections. These sections behave as harmonic oscillators. We assume that we have $N + 1$ oscillators and therefore we divide the interval $[0, L]$ into $N + 1$ equidistant subintervals (with length ΔX). The oscillator at $n = 0$ is part of the middle ear, the oscillator at $n = 1$ has the highest frequency in the cochlea and the oscillator at $n = N$, at the right the helicotrema, has the lowest frequency in the cochlea.

The following approximation of the second partial derivative of $p(x, t)$ [Vuik et al. 2006] is used:

$$\frac{\partial^2 p}{\partial x^2}(x, t) \approx \frac{p(x + \Delta X, t) - 2p(x, t) + p(x - \Delta X, t)}{(\Delta X)^2}.$$

Then (2) can be written as a matrix representation $\mathbf{A}\mathbf{p}(t) = \mathbf{b}(t)$ [Vuik et al. 2006]. Here matrix \mathbf{A} is a tridiagonal matrix (a_{ij}) , vector $\mathbf{p}(t)$ is an unknown vector and the vector $\mathbf{b}(t)$ consists of known terms like the stimulus $p_e(t)$ and the vector $\mathbf{G}(t)$:

$$a_{ij} = \begin{cases} (1 + \frac{2m_{c01}}{m_m})/\Delta X & \text{if } i = j = 1, \\ (2 + \frac{2m_c}{m})/\Delta X & \text{if } i = j \text{ and } i \in \{2, 3, \dots, N\}, \\ (1 + \frac{2m_c}{m} + \frac{2m_c}{m_h + 2m_c})/\Delta X & \text{if } i = j = N + 1, \\ -1/\Delta X & \text{if } j = i + 1 \text{ and } i \in \{1, 2, \dots, N\}, \\ -1/\Delta X & \text{if } j = i - 1 \text{ and } i \in \{2, 3, \dots, N + 1\}, \end{cases}$$

$$p_i = p(x_i, t) \quad \text{for } i \in \{0, 1, \dots, N\},$$

$$b_j = \begin{cases} \frac{2m_{c01}}{m_m}(n_t p_e(t) + G(x_{j-1}, t))/\Delta X & \text{if } j = 1, \\ \frac{2m_c}{m}G(x_{j-1}, t)/\Delta X & \text{if } j \in \{2, 3, \dots, N + 1\}. \end{cases}$$

This almost corresponds with the system derived by Diependaal et al., but the equation associated with the oscillator at $n = 0$ (first equation of $\mathbf{A}\mathbf{p}(t) = \mathbf{b}(t)$) differs. This deviation is a result of the fact that the oscillator at $n = 0$ is part of the middle ear. The other equations do correspond with that of Diependaal et al. for equidistant subintervals (ΔX is constant), except the notation deviates:

$$\frac{2m_c}{m} = \frac{2\rho b_{BM}}{m_s S_{sc}}(\Delta X)^2 \quad \text{and} \quad \frac{2m_c}{m_h + 2m_c} \approx 0.999999999102402 \approx 1.$$

2.4. System of equations for the time. Consider the functions $\dot{y}(x, t)$ and $\ddot{y}(x, t)$. These lead to a second-order system of time equations. This second-order system can be transformed into a first-order system if we define $\dot{y}(x, t) = u(x, t)$. Then it holds that $\ddot{y}(x, t) = \dot{u}(x, t)$ [Diependaal et al. 1987] and $\dot{u}(x, t) = \ddot{y}(x, t) = (p(x, t) - G(x, t))/m_s$.

So this system is given by

$$\begin{cases} \dot{y}(x, t) = u(x, t), \\ \dot{u}(x, t) = (p(x, t) - G(x, t))/m_s, \end{cases} \quad \begin{cases} y(x, 0) = 0, & 0 \leq x \leq L, \\ u(x, 0) = 0, & t \geq 0. \end{cases}$$

It is transformed into a system consisting of vectors after spatial discretisation has taken place because the vector $\mathbf{p}(t)$ can be determined from $\mathbf{A}\mathbf{p}(t) = \mathbf{b}(t)$.

Consider the system

$$\begin{cases} \dot{\mathbf{y}}(t) = \mathbf{u}(t), \\ \dot{\mathbf{u}}(t) = \mathbf{Q} \cdot [\mathbf{p}(t) - \mathbf{c}(t)], \end{cases} \quad t \geq 0, \quad \begin{cases} \mathbf{y}(0) = \mathbf{0}, \\ \mathbf{u}(0) = \mathbf{0}, \end{cases} \quad (3)$$

where the matrix \mathbf{Q} is a diagonal matrix (q_{ii}) and the vector $\mathbf{c}(t)$ consists of known terms like the stimulus $p_e(t)$ and the vector $\mathbf{G}(t)$:

$$q_{ii} = \begin{cases} 1/m_{sm} & \text{if } i = 1, \\ 1/m_s & \text{if } i \in \{2, 3, \dots, N+1\}, \end{cases}$$

$$c_j = \begin{cases} n_t p_e(t) + G(x_{j-1}, t) & \text{if } j = 1, \\ G(x_{j-1}, t) & \text{if } j \in \{2, 3, \dots, N+1\}. \end{cases}$$

Because of the oscillator in the middle ear, this system deviates a little bit for $n = 0$ (first equation of $\mathbf{Q}[\mathbf{p}(t) - \mathbf{c}(t)]$) with respect to [Diependaal et al. 1987]. However, in this case the influence of the middle ear is taken into account.

The vector $\mathbf{p}(t)$ is determined by

$$\mathbf{p}(t) = \mathbf{A}^{-1} \mathbf{b}(t) = \mathbf{A}^{-1} \begin{pmatrix} -\frac{2m_{c01}}{m_m} G(x_0, t) \\ -\frac{2m_c}{m} G(x_1, t) \\ \vdots \\ -\frac{2m_c}{m} G(x_N, t) \end{pmatrix} - \mathbf{A}^{-1} \begin{pmatrix} -\frac{2m_{c01}}{m_m} n_t p_e(t) \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

3. Numerical methods

3.1. System of first-order differential equations. We consider system (3). In the following definition of the vector $\mathbf{G}(t)$, it can be seen that the properties of the middle ear are again taken into account for the first oscillator, and therefore it differs from the vector $\mathbf{g}(t)$ in the paper of Diependaal et al.:

$$\mathbf{G}(t) = \begin{pmatrix} G(x_0, t) \\ G(x_1, t) \\ \vdots \\ G(x_N, t) \end{pmatrix} = \begin{pmatrix} S_{ST} \cdot n_t^2 \cdot Z_a \cdot u(x_0, t) + s_{sm} \cdot y(x_0, t) \\ d_s(x_1) \cdot u(x_1, t) + s_s(x_1) \cdot y(x_1, t) \\ \vdots \\ d_s(x_N) \cdot u(x_N, t) + s_s(x_N) \cdot y(x_N, t) \end{pmatrix} = \mathbf{D}\mathbf{u}(t) + \mathbf{S}\mathbf{y}(t).$$

However, for the oscillators in the cochlea, $\mathbf{G}(t)$ equals $\mathbf{g}(t)$, but again a somewhat different notation is used. Here the matrices \mathbf{D} and \mathbf{S} are diagonal matrices (d_{ii}) and (s_{ii}) respectively given by

$$d_{ii} = \begin{cases} S_{ST} \cdot n_t^2 \cdot Z_a & \text{if } i = 1, \\ d_s(x_{i-1}) & \text{if } i \in \{2, 3, \dots, N+1\}, \end{cases}$$

$$s_{ii} = \begin{cases} s_{sm} & \text{if } i = 1, \\ s_s(x_{i-1}) & \text{if } i \in \{2, 3, \dots, N+1\}. \end{cases}$$

Consider system (3) on the time interval $[0, T]$. This time interval is divided into M equidistant subintervals $[t_0, t_1], [t_1, t_2], \dots, [t_{M-1}, t_M]$ (with length Δt). After dividing the time interval into subintervals, the following steps must be followed from $j = 1$ to $j = M$ [Diependaal et al. 1987]:

- (1) Calculate at time t_{j-1} the vectors \mathbf{c} and \mathbf{b} .
 - (2) Solve \mathbf{p} by using $\mathbf{A}\mathbf{p}(t) = \mathbf{b}(t)$.
 - (3) Calculate $\mathbf{w}[t, \mathbf{y}(t), \mathbf{u}(t)] = \mathbf{Q} \cdot [\mathbf{p}(t) - \mathbf{c}(t)]$.
 - (4) Integrate the equations $\dot{\mathbf{y}}(t) = \mathbf{u}(t)$ and $\dot{\mathbf{u}}(t) = \mathbf{w}[t, \mathbf{y}(t), \mathbf{u}(t)]$ from t_{j-1} to t_j .
- At step (4), the fourth-order Runge–Kutta and modified Sielecki methods are used.

3.2. Fourth-order Runge–Kutta method. The fourth-order Runge–Kutta method is given in [Diependaal et al. 1987] by

$$\begin{cases} \mathbf{y}(t + \Delta t) = \mathbf{y}(t) + \frac{1}{6}[\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4], \\ \mathbf{u}(t + \Delta t) = \mathbf{u}(t) + \frac{1}{6}[\mathbf{l}_1 + 2\mathbf{l}_2 + 2\mathbf{l}_3 + \mathbf{l}_4]. \end{cases}$$

The following four steps are carried out:

- (1) Determine the predictors \mathbf{k}_1 and \mathbf{l}_1 .
- (2) Determine the predictors \mathbf{k}_2 and \mathbf{l}_2 .
- (3) Determine the predictors \mathbf{k}_3 and \mathbf{l}_3 .
- (4) Determine the predictors \mathbf{k}_4 and \mathbf{l}_4 .

At each step the vector $\mathbf{p}(t)$ has to be determined from $\mathbf{A}\mathbf{p}(t) = \mathbf{b}(t)$ for the function $\mathbf{w}[t, \mathbf{y}(t), \mathbf{u}(t)] = \mathbf{Q} \cdot [\mathbf{p}(t) - \mathbf{c}(t)]$.

To apply the fourth-order Runge–Kutta method, system (3) must be written as the matrix representation

$$\begin{pmatrix} \dot{\mathbf{y}}(x_0, t) \\ \dot{\mathbf{y}}(x_1, t) \\ \vdots \\ \dot{\mathbf{y}}(x_N, t) \\ \dot{\mathbf{u}}(x_0, t) \\ \dot{\mathbf{u}}(x_1, t) \\ \vdots \\ \dot{\mathbf{u}}(x_N, t) \end{pmatrix} = \begin{pmatrix} \mathbf{O} & \mathbf{I} \\ \mathbf{M}\mathbf{S} & \mathbf{M}\mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{y}(x_0, t) \\ \mathbf{y}(x_1, t) \\ \vdots \\ \mathbf{y}(x_N, t) \\ \mathbf{u}(x_0, t) \\ \mathbf{u}(x_1, t) \\ \vdots \\ \mathbf{u}(x_N, t) \end{pmatrix} - \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{N} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ n_t \mathbf{p}_e(t) \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \quad (4)$$

The matrix \mathbf{O} is an $(N+1) \times (N+1)$ null matrix, \mathbf{I} is an $(N+1) \times (N+1)$ identity matrix, \mathbf{M} is determined by $\mathbf{M} = \mathbf{Q}[\mathbf{A}^{-1}\mathbf{R} - \mathbf{I}]$ with \mathbf{R} an $(N+1) \times (N+1)$ diagonal matrix with

$$r_{ii} = \begin{cases} -\frac{2m_{c01}}{m_m} & \text{if } i = 1, \\ -\frac{2m_c}{m} & \text{if } i \in \{2, 3, \dots, N+1\}, \end{cases}$$

and \mathbf{N} is an $(N+1) \times (N+1)$ matrix determined by $\mathbf{N} = \mathbf{Q}\left[-\frac{2m_{c01}}{m_m} \cdot \mathbf{A}^{-1} + \mathbf{I}\right]$.

3.3. The properties of the fourth-order Runge–Kutta method. The Runge–Kutta method we are using is a fourth-order method, which means that the total error is of the fourth-order (with respect to the time step). This is an explicit method and it is conditionally stable, which means that it is stable for a time step below a certain bound.

In order to study the stability when using the time-integration fourth-order Runge–Kutta method, the amplification factor associated with this method can be used and it is given by

$$Q(h\lambda) = 1 + h\lambda + \frac{1}{2}(h\lambda)^2 + \frac{1}{6}(h\lambda)^3 + \frac{1}{24}(h\lambda)^4,$$

where h represents the time step used and λ is known from the test equation $y' = \lambda y$. A numerical time-integration method is termed stable if and only if $|Q(h\lambda)| \leq 1$ for a given time step h [Vuik et al. 2006].

The amplification factor $Q(h\lambda)$ given above is the amplification factor in the scalar case, but here the method is used on a matrix representation (4). Now this numerical time-integration method is termed stable if and only if for every eigenvalue μ of the matrix $\begin{pmatrix} O & I \\ MS & MD \end{pmatrix}$, it holds that $|Q(h\mu)| \leq 1$ for a given time step h [Vuik et al. 2006].

3.4. The modified Sielecki method. The modified Sielecki method is given in [Diependaal et al. 1987] by

$$\begin{cases} \mathbf{u}(t + \Delta t) = \mathbf{u}(t) + \Delta t \cdot \mathbf{w}[t, \mathbf{y}(t), \mathbf{u}(t)], \\ \mathbf{y}(t + \Delta t) = \mathbf{y}(t) + \Delta t \cdot \mathbf{u}(t + \Delta t), \end{cases}$$

where the function $\mathbf{w}[t, \mathbf{y}(t), \mathbf{u}(t)]$ is defined as $\mathbf{w}[t, \mathbf{y}(t), \mathbf{u}(t)] = \mathbf{Q} \cdot [\mathbf{p}(t) - \mathbf{c}(t)]$, which can also be represented as

$$\mathbf{w}[t, \mathbf{y}(t), \mathbf{u}(t)] = \mathbf{M}\mathbf{D}\mathbf{u}(t) + \mathbf{M}\mathbf{S}\mathbf{y}(t) - \mathbf{N} \begin{pmatrix} n_t p_e(t) \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

3.5. The stability of the modified Sielecki method. The amplification factor of the modified Sielecki numerical time-integration method is not known or given as far as the authors know. This method is both an implicit and an explicit method and this makes it harder to derive an amplification factor.

To examine the stability of the modified Sielecki method, the scalar-linear case is first considered. For the (scalar) system

$$\begin{cases} \dot{\zeta}(t) = \nu(t), & \zeta(0) = \zeta_0, \\ \dot{\nu}(t) = \omega[t, \zeta(t), \nu(t)], & \nu(0) = \nu_0, \end{cases}$$

the modified Sielecki method is given by

$$\begin{cases} v(t + \Delta t) = v(t) + \Delta t \cdot \omega[t, \zeta(t), v(t)], \\ \zeta(t + \Delta t) = \zeta(t) + \Delta t \cdot v(t + \Delta t). \end{cases}$$

In the scalar-linear case the function $\omega[t, \zeta(t), v(t)]$ is given by $\omega[t, \zeta(t), v(t)] = \lambda \cdot \zeta(t) + \mu \cdot v(t) + c$ with $\lambda, \mu \leq 0$ and a constant $c \in \mathbb{R}$. To examine the stability of this method in the scalar-linear case, the function $\omega[t, \zeta(t), v(t)] = \lambda \cdot \zeta(t) + \mu \cdot v(t)$ can be considered because the constant c does not affect the stability.

Consider the (scalar-linear) system

$$\begin{cases} v(t + \Delta t) = v(t) + \Delta t \cdot \lambda \cdot \zeta(t) + \Delta t \cdot \mu \cdot v(t), \\ \zeta(t + \Delta t) = \zeta(t) + \Delta t \cdot v(t + \Delta t). \end{cases} \quad (5)$$

After substituting $v(t + \Delta t) = v(t) + \Delta t \cdot \lambda \cdot \zeta(t) + \Delta t \cdot \mu \cdot v(t)$ in $\zeta(t + \Delta t) = \zeta(t) + \Delta t \cdot v(t + \Delta t)$, the system (5) can be represented by the matrix representation $\mathbf{a}(t + \Delta t) = \mathbf{T} \cdot \mathbf{a}(t)$. Here

$$\mathbf{a}(t + \Delta t) = \begin{pmatrix} v(t + \Delta t) \\ \zeta(t + \Delta t) \end{pmatrix}, \quad \mathbf{a}(t) = \begin{pmatrix} v(t) \\ \zeta(t) \end{pmatrix}, \quad \mathbf{T} = \begin{pmatrix} 1 + \Delta t \cdot \mu & \Delta t \cdot \lambda \\ \Delta t + (\Delta t)^2 \cdot \mu & 1 + (\Delta t)^2 \cdot \lambda \end{pmatrix}.$$

For a multiplicative norm, it holds for $\mathbf{a}_{n+1} = \mathbf{T} \cdot \mathbf{a}_n$ with \mathbf{a}_n the numerical solution that

$$\|\mathbf{a}_{n+1}\| = \|\mathbf{T} \cdot \mathbf{a}_n\| = \|\mathbf{T}^n \cdot \mathbf{a}_1\| \stackrel{\text{multiplicativity}}{\leq} \|\mathbf{T}^n\| \cdot \|\mathbf{a}_1\|.$$

Furthermore the following result is known [Golub and Van Loan 1996]:

$$\mathbf{T}^n \rightarrow 0 \quad \text{as } \rho(\mathbf{T}) < 1 \quad \text{with } \rho(\mathbf{T}) = \max\{|\kappa| : \kappa \text{ is an eigenvalue of } \mathbf{T}\}.$$

The modified Sielecki numerical time-integration method is termed stable if and only if $|\kappa| \leq 1$ for every eigenvalue κ of \mathbf{T} for a given time step Δt , because then the inequality $\rho(\mathbf{T}) < 1$ is satisfied.

The same principle (as in the scalar-linear case) can be used for

$$\begin{cases} \mathbf{u}(t + \Delta t) = \mathbf{u}(t) + \Delta t \cdot \mathbf{w}[t, \mathbf{y}(t), \mathbf{u}(t)], \\ \mathbf{y}(t + \Delta t) = \mathbf{y}(t) + \Delta t \cdot \mathbf{u}(t + \Delta t), \end{cases}$$

where the function $\mathbf{w}[t, \mathbf{y}(t), \mathbf{u}(t)] = \mathbf{M}\mathbf{D}\mathbf{u}(t) + \mathbf{M}\mathbf{S}\mathbf{y}(t)$ is considered to examine the stability of the modified Sielecki method. So consider

$$\begin{cases} \mathbf{u}(t + \Delta t) = \mathbf{u}(t) + \Delta t \cdot \mathbf{M}\mathbf{S}\mathbf{y}(t) + \Delta t \cdot \mathbf{M}\mathbf{D}\mathbf{u}(t), \\ \mathbf{y}(t + \Delta t) = \mathbf{y}(t) + \Delta t \cdot \mathbf{u}(t + \Delta t). \end{cases} \quad (6)$$

After substituting $\mathbf{u}(t + \Delta t) = \mathbf{u}(t) + \Delta t \cdot \mathbf{MSy}(t) + \Delta t \cdot \mathbf{MDu}(t)$ in $\mathbf{y}(t + \Delta t) = \mathbf{y}(t) + \Delta t \cdot \mathbf{u}(t + \Delta t)$, the system (6) can be written as

$$\begin{pmatrix} \mathbf{u}(t + \Delta t) \\ \mathbf{y}(t + \Delta t) \end{pmatrix} = \begin{pmatrix} \mathbf{I} + \Delta t \cdot \mathbf{MD} & \Delta t \cdot \mathbf{MS} \\ \Delta t \cdot \mathbf{I} + (\Delta t)^2 \cdot \mathbf{MD} & \mathbf{I} + (\Delta t)^2 \cdot \mathbf{MS} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{u}(t) \\ \mathbf{y}(t) \end{pmatrix}. \quad (7)$$

The modified Sielecki method is stable if and only if $|\kappa| \leq 1$ holds for every eigenvalue κ of the matrix

$$\begin{pmatrix} \mathbf{I} + \Delta t \cdot \mathbf{MD} & \Delta t \cdot \mathbf{MS} \\ \Delta t \cdot \mathbf{I} + (\Delta t)^2 \cdot \mathbf{MD} & \mathbf{I} + (\Delta t)^2 \cdot \mathbf{MS} \end{pmatrix}$$

for a given time step Δt .

4. Numerical experiments

4.1. Problem. The one-dimensional cochlea model is given by (1) and this equation can be written as a matrix representation (spatial discretisation) $\mathbf{Ap}(t) = \mathbf{b}(t)$ [Vuik et al. 2006], which is used to determine the vector $\mathbf{p}(t)$. The system of first-order time equations (3) can be solved by determining the vector $\mathbf{p}(t)$ and using a numerical time-integration method. In this paper the fourth-order Runge–Kutta and modified Sielecki numerical time-integration methods are dealt with.

4.2. The stability of the two numerical time-integration methods. For the fourth-order Runge–Kutta method, system (3) is written as the matrix representation (4).

The fourth-order Runge–Kutta method is termed stable if and only if for every eigenvalue μ_i of the matrix

$$\mathbf{E} = \begin{pmatrix} \mathbf{O} & \mathbf{I} \\ \mathbf{MS} & \mathbf{MD} \end{pmatrix},$$

it holds that $|Q(\Delta t \mu_i)| \leq 1$ for a given time step Δt [Vuik et al. 2006], where the amplification factor Q is given by

$$Q(\Delta t \mu_i) = 1 + \Delta t \mu_i + \frac{1}{2}(\Delta t \mu_i)^2 + \frac{1}{6}(\Delta t \mu_i)^3 + \frac{1}{24}(\Delta t \mu_i)^4.$$

For the modified Sielecki method, $\mathbf{w}[t, \mathbf{y}(t), \mathbf{u}(t)] = \mathbf{MDu}(t) + \mathbf{MSy}(t)$ is considered to examine the stability. Consider the system (6) and write this as the matrix representation given in (7).

The modified Sielecki method will be stable if and only if $|\kappa_j| \leq 1$ holds for every eigenvalue κ_j of the matrix

$$\mathbf{F} = \begin{pmatrix} \mathbf{I} + \Delta t \cdot \mathbf{MD} & \Delta t \cdot \mathbf{MS} \\ \Delta t \cdot \mathbf{I} + (\Delta t)^2 \cdot \mathbf{MD} & \mathbf{I} + (\Delta t)^2 \cdot \mathbf{MS} \end{pmatrix}$$

for a given time step Δt .

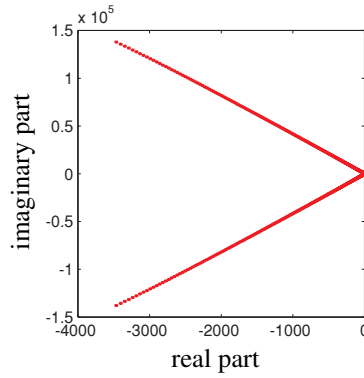


Figure 2. The plot of the eigenvalues of E .

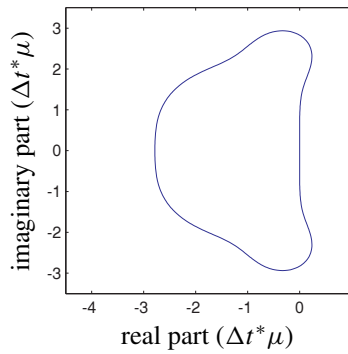


Figure 3. The stability locus of the fourth-order Runge–Kutta method.

4.3. Numerical experiments to determine a restriction on the time step Δt . To perform these numerical experiments, Matlab is used.

4.3.1. Fourth-order Runge–Kutta. The eigenvalues μ_i of the matrix E can be calculated and plotted (see Figure 2).

For the fourth-order Runge–Kutta method to be stable for a given Δt , the inequality

$$|Q(\Delta t \mu_i)| = \left| 1 + \Delta t \mu_i + \frac{1}{2}(\Delta t \mu_i)^2 + \frac{1}{6}(\Delta t \mu_i)^3 + \frac{1}{24}(\Delta t \mu_i)^4 \right| \leq 1$$

has to be satisfied for all μ_i ($i = 1, \dots, 2N + 2$). In other words, all eigenvalues μ_i multiplied by a time step Δt must lie within the range of the stability locus of the fourth-order Runge–Kutta method. Figure 3 shows this stability locus.

This consideration determines a restriction on the time step Δt . The result of this numerical experiment is the following:

- For $\Delta t = 2.08 \cdot 10^{-5} s$, it holds that $|Q(\Delta t \mu_i)| \leq 1$ for all eigenvalues μ_i , and thus this numerical scheme is stable.

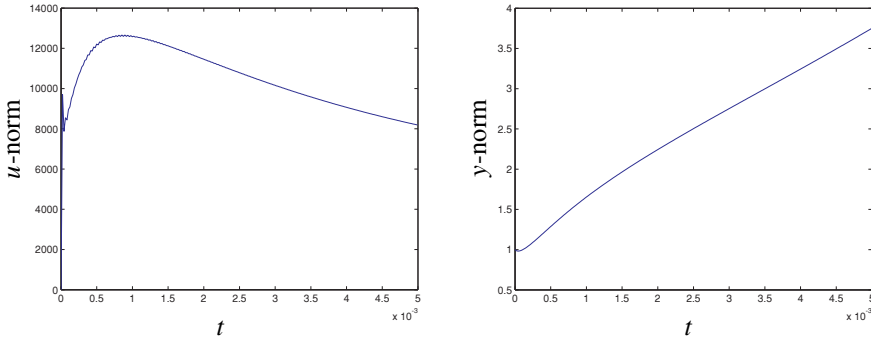


Figure 4. The norms of $\mathbf{u}(t)$ and $\mathbf{y}(t)$, divided by $\sqrt{N+1}$, with $\Delta t = 2.08 \cdot 10^{-5} s$.

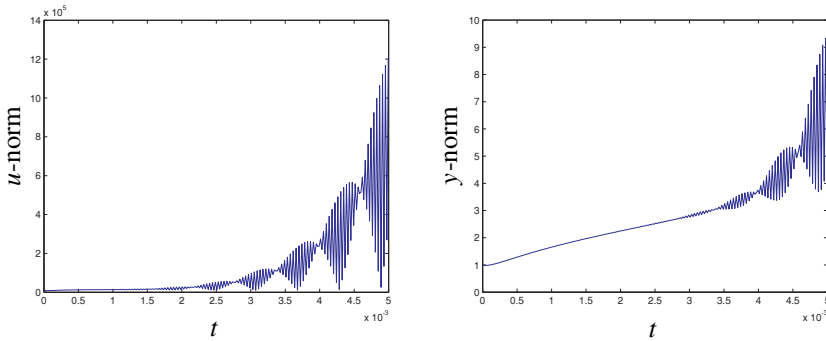


Figure 5. The norms of $\mathbf{u}(t)$ and $\mathbf{y}(t)$, divided by $\sqrt{N+1}$, with $\Delta t = 2.09 \cdot 10^{-5} s$.

- For $\Delta t = 2.09 \cdot 10^{-5} s$ it holds that $|Q(\Delta t \mu_i)| > 1$ for two eigenvalues, and thus this numerical scheme is no longer stable.

The determined time steps ($\Delta t = 2.08 \cdot 10^{-5} s$ and $\Delta t = 2.09 \cdot 10^{-5} s$) can also be tested on the fourth-order Runge–Kutta numerical time-integration method. The initial conditions $\mathbf{u}(0) = \mathbf{1}$, $\mathbf{y}(0) = \mathbf{1}$ are used instead of $\mathbf{u}(0) = \mathbf{0}$, $\mathbf{y}(0) = \mathbf{0}$. A small change of the initial conditions causes a perturbation.

If this perturbation is bounded, then the fourth-order Runge–Kutta method is stable. The expectation is that the perturbation is bounded by using the time step $\Delta t = 2.08 \cdot 10^{-5} s$ and unbounded by using the time step $\Delta t = 2.09 \cdot 10^{-5} s$. To investigate this, the norms of the vectors $\mathbf{u}(t)$ and $\mathbf{y}(t)$ divided by $\sqrt{N+1}$ are calculated and plotted (see Figure 4 for time step $\Delta t = 2.08 \cdot 10^{-5} s$ and Figure 5 for time step $\Delta t = 2.09 \cdot 10^{-5} s$).

From Figure 4, it can be seen that the perturbation remains bounded. For $\Delta t = 2.08 \cdot 10^{-5} s$, the numerical scheme is indeed stable.

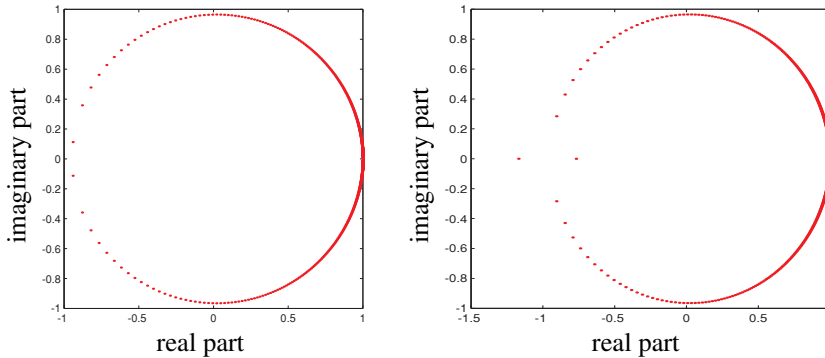


Figure 6. The eigenvalues of F with $\Delta t = 1.41 \cdot 10^{-5}s$ (left) and $\Delta t = 1.42 \cdot 10^{-5}s$ (right).

From Figure 5, it can be concluded that the perturbation is unbounded. For $\Delta t = 2.09 \cdot 10^{-5}s$, it holds that the numerical scheme is unstable.

4.3.2. Modified Sielecki. By trying different values for the time step Δt , we can determine for which time step the modified Sielecki numerical time-integration method is stable or unstable. It can be seen that for the time step $\Delta t = 1.41 \cdot 10^{-5}s$, the inequality $|\kappa_j| < 1$ holds for all eigenvalues κ_j of the matrix F (see Figure 6). It can also be seen that for the time step $\Delta t = 1.42 \cdot 10^{-5}s$, the inequality $|\kappa_j| > 1$ holds for one eigenvalue (see Figure 6).

The determined time steps ($\Delta t = 1.41 \cdot 10^{-5}s$ and $\Delta t = 1.42 \cdot 10^{-5}s$) can also be tested on the modified Sielecki numerical time-integration method. The initial conditions $\mathbf{u}(0) = \mathbf{1}$, $\mathbf{y}(0) = \mathbf{1}$ instead of $\mathbf{u}(0) = \mathbf{0}$, $\mathbf{y}(0) = \mathbf{0}$ are used, causing a perturbation.

If this perturbation is bounded, then the modified Sielecki method is stable. The expectation is that the perturbation is bounded by using the time step $\Delta t = 1.41 \cdot 10^{-5}s$ and unbounded by using the time step $\Delta t = 1.42 \cdot 10^{-5}s$. Again the norms of the vectors $\mathbf{u}(t)$ and $\mathbf{y}(t)$ divided by $\sqrt{N+1}$ are calculated and plotted (see Figure 7 for time step $\Delta t = 1.41 \cdot 10^{-5}s$ and Figure 8 for time step $\Delta t = 1.42 \cdot 10^{-5}s$).

From Figure 7, it can be seen that the perturbation remains bounded. For $\Delta t = 1.41 \cdot 10^{-5}s$ the numerical scheme is indeed stable.

From Figure 8, it can be concluded that the perturbation is unbounded. For $\Delta t = 1.42 \cdot 10^{-5}s$ it holds that the numerical scheme is unstable.

5. Conclusions

After examining the stability of the fourth-order Runge–Kutta and modified Sielecki numerical time-integration methods, it can be concluded that a bigger time step

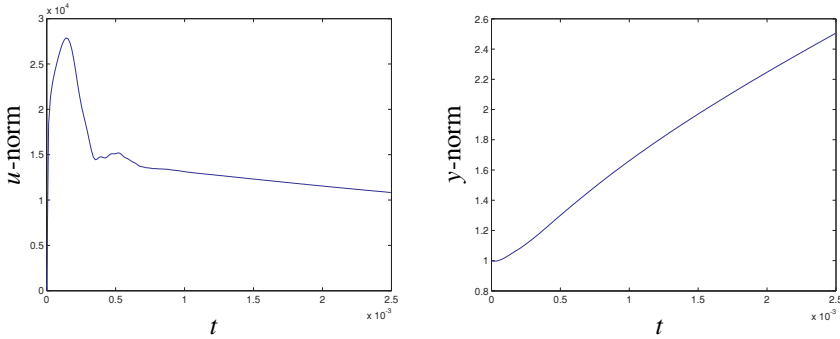


Figure 7. The norms of $\mathbf{u}(t)$ and $\mathbf{y}(t)$, divided by $\sqrt{N+1}$, with $\Delta t = 1.41 \cdot 10^{-5} s$.

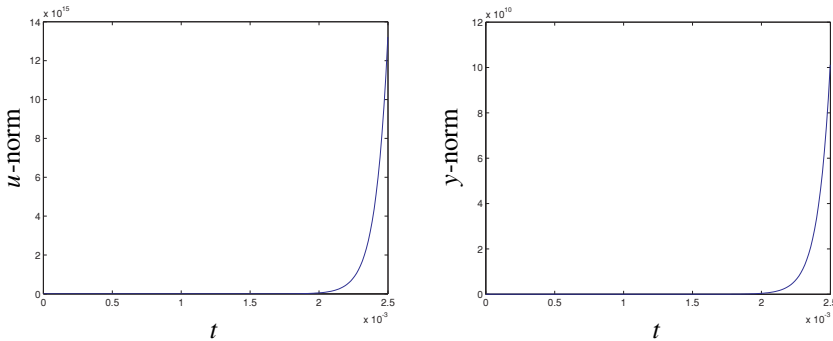


Figure 8. The norms of $\mathbf{u}(t)$ and $\mathbf{y}(t)$, divided by $\sqrt{N+1}$, with $\Delta t = 1.42 \cdot 10^{-5} s$.

($\Delta t = 2.08 \cdot 10^{-5} s$) can be used for the Runge–Kutta four method than for the method modified Sielecki ($\Delta t = 1.41 \cdot 10^{-5} s$). This corresponds with the article by Diependaal, Duifhuis, Hoogstraten and Viergever [Diependaal et al. 1987].

When the time step $\Delta t = 2.08 \cdot 10^{-5} s$ is used, the fourth-order Runge–Kutta method is still stable, but the modified Sielecki method is then unstable. The modified Sielecki method already shows unstable behavior when a time step of $\Delta t = 1.42 \cdot 10^{-5} s$ is used.

Our numerical stability analysis for both time integration methods showed the following results:

- *Fourth-order Runge–Kutta:* when we use a time step of $2.08 \cdot 10^{-5} s$, the system is numerically stable, but numerically unstable for a time step of $2.09 \cdot 10^{-5} s$.
- *Modified Sielecki:* a time step of $1.41 \cdot 10^{-5} s$ causes the system to be numerically stable, and a time step of $1.42 \cdot 10^{-5} s$ causes it to be numerically unstable.

Following this numerical stability analysis, we tried to verify these results with a numerical test (Figure 4, 5, 7 and 8), and saw that the numerical tests supported the results from our analysis. Thus, the theoretical analyses and experimental analyses coincide.

List of symbols

- $p(x, t)$: transmembrane fluid pressure
- $y(x, t)$: excitation of the basilar membrane
- m_s : specific acoustic mass of the basilar membrane ($m_s = m \cdot b \cdot \Delta X$)
- $d_s(x)$: specific acoustic damping of the basilar membrane at place x ($d_s(x) = d(x) \cdot b \cdot \Delta X$ and $d_s(x_n) = d_{sAMP}(x_n) \cdot d_{sPROF}(x_n)$)
- $d_{sAMP}(x)$: ensures that the damping in the cochlea is uniform everywhere ($d_{sAMP}(x) = \epsilon \sqrt{m_s s_s(x)}$)
- $d_{sPROF}(x)$: makes it possible to locally vary the (negative) damping ($d_{sPROF}(x) = 1$ in the linear case)
- ϵ : models strength impulse response (Matlab: $\epsilon = 5 \cdot 10^{-2}$)
- $s_s(x)$: specific acoustic stiffness of the basilar membrane at place x ($s_s(x) = s(x) \cdot b \cdot \Delta X$ and $s_s(x_n) = s_0 \cdot e^{-\lambda x_n}$)
- s_0 : specific acoustic stiffness constant (Matlab: $s_0 = 1 \cdot 10^{10}$ Pa/m)
- λ : value which determines place-frequency relation in the cochlea (Matlab: $\lambda = 300 \text{ m}^{-1}$)
- ρ : density of the cochlear fluid
- h : height of a scala ($h = S_{sc}/b$)
- m : acoustic mass of the basilar membrane
- $d(x)$: acoustic damping of the basilar membrane at place x
- $s(x)$: acoustic stiffness of the basilar membrane at place x
- ΔX : length of a subinterval ($\Delta X = L/(N + 1)$)
- S_{sc} : surface of a scala
- b : width of a scala
- m_{c01} : acoustic mass of the cochlear fluid between the oval window and the first oscillator ($m_{c01} = \rho \Delta X_{01}/S_{sc}$)
- ΔX_{01} : distance between the oval window and the first oscillator
- m_m : acoustic mass of the middle ear ($m_m = n_t^2 Z_s / (S_t \omega_{rm} \delta)$)
- m_{sm} : specific acoustic mass of the middle ear ($m_{sm} = m_m S_{ST}$)
- Z_s : specific acoustic impedance of air
- S_t : surface of the eardrum

- ω_{rm} : resonance frequency of the middle ear
- δ : reciprocal value of the quality factor Q of the middle ear ($\delta = d_m / \sqrt{s_m m_m}$)
- d_m : acoustic damping of the middle ear
- s_m : acoustic stiffness of the middle ear
- n_t : transformation factor of the middle ear
- $p_e(t)$: form of the stimulus
- m_c : acoustic mass of the cochlear fluid between two oscillators ($m_c = \rho \Delta X / S_{sc}$)
- $u(x, t)$: velocity of the basilar membrane ($u(x, t) = \dot{y}(x, t)$)
- S_{ST} : surface of the stapes (Matlab: $S_{ST} = 3 \cdot 10^{-6} \text{ m}^2$)
- Z_a : acoustic impedance of air ($Z_a = Z_s / S_t$)
- s_{sm} : specific acoustic stiffness of the middle ear
- Δt : time step used in the fourth-order Runge–Kutta and modified Sielecki numerical methods

References

- [Bell 2004] A. Bell, “Hearing: travelling wave or resonance?”, *PLoS Biol* **2**:10 (2004), Art. ID e337.
- [Diependaal et al. 1987] R. J. Diependaal, H. Duifhuis, H. W. Hoogstraten, and M. A. Viergever, “Numerical methods for solving one-dimensional cochlear models in the time domain”, *J. Acoust. Soc. Am.* **82**:5 (1987), 1655–1666.
- [Golub and Van Loan 1996] G. H. Golub and C. F. Van Loan, *Matrix computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996. MR Zbl
- [van Hengel 1996] P. W. J. van Hengel, *Emissions from cochlear modelling*, Ph.D. thesis, University of Gronigen, 1996, <http://hdl.handle.net/11370/2c233b60-f37e-440d-b304-5aaa53c6605c>.
- [Robles and Ruggero 2001] L. Robles and M. A. Ruggero, “Mechanics of the mammalian cochlea”, *Physiological Reviews* **81**:3 (2001), 1305–1352.
- [Vuik et al. 2006] C. Vuik, P. van Beek, F. Vermolen, and J. en van Kan, *Numerical methods for ordinary differential equations*, 1st ed., Delft Academic Press/VSSD, 2006.

Received: 2014-05-09 Revised: 2015-04-14 Accepted: 2016-03-25

kimberleylindenberg@gmail.com *Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science, Department of Applied Mathematical Analysis, 2628 CD Delft, The Netherlands*

INCAS³, 9401 HJ Assen, The Netherlands

c.vuik@tudelft.nl *Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science, Department of Applied Mathematical Analysis, 2628 CD Delft, The Netherlands*

petervanhengel@incas3.eu *INCAS³, 9401 HJ Assen, The Netherlands*

Three approaches to a bracket polynomial for singular links

Carmen Caprau, Alex Chichester and Patrick Chu

(Communicated by Jim Hoste)

In this paper we extend the Kauffman bracket to singular links. Specifically, we define a polynomial invariant for singular links, and in doing this, we consider three approaches to our extended Kauffman bracket polynomial: (1) using skein relations involving singular link diagrams, (2) using representations of the singular braid monoid, (3) via a Yang–Baxter state model. We also study some properties of the extended Kauffman bracket.

1. Introduction and background

Knot theory is one of the most active research areas in mathematics. In the recent years, there has been a great interest in the study of knot-like objects, including singular links, knotted graphs, virtual knots and pseudoknots, not only because of their connections to other areas in mathematics, but also because of their applications to physics, chemistry, and molecular biology.

In this paper we focus on singular links and construct an invariant for such objects, based on the skein relation defining the Kauffman bracket for classical knots and links. We hope our work will prove useful for young researchers interested in knot theory for its intrinsic beauty or for its possible applications.

Knot theory studies embeddings of circles in three-dimensional space. When more than one circle is embedded in \mathbb{R}^3 , the resulting embedding is called a *link*; otherwise, it is called a *knot*. In particular, a link is a disjoint union of knots, and these knots are called the components of the link. For simplicity, whenever possible, we will refer to both knots and links as knots. A diagram of a knot is a projection of the knot into a plane, and the *crossings* of a knot diagram are artifacts of the projection. We consider only *regular diagrams*, in which all crossings are double points.

A *singular link* is an immersion of a disjoint union of circles in three-dimensional space, which has finitely many singularities, called *singular crossings*, that are all

MSC2010: 57M25, 57M27.

Keywords: Kauffman bracket, invariants for knots and links, singular braids and links, Yang–Baxter equation.

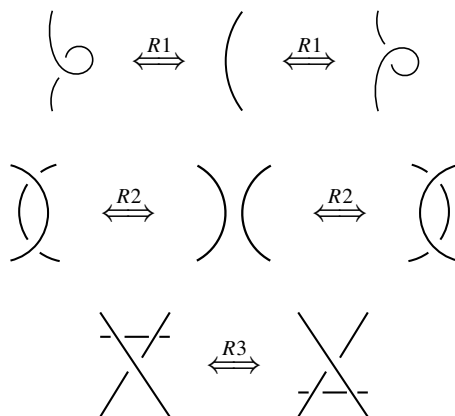


Figure 1. The Reidemeister moves.

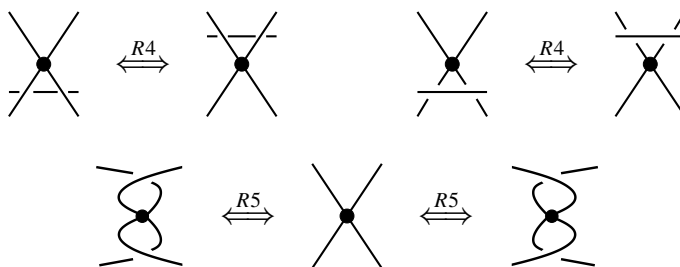


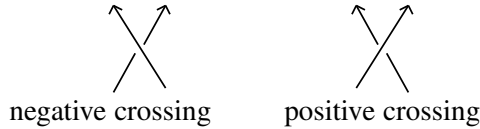
Figure 2. Additional moves for singular links.

transverse double points. A singular link can be regarded as an embedding in \mathbb{R}^3 of a four-valent graph with *rigid vertices*. We can think of such vertices as being rigid disks with four strands connected to it which turn as a whole when we flip the vertex by 180 degrees.

The goal of knot theory is to know whether or not two knots are isotopic. Two knots are called *ambient isotopic* if there is a continuously varying family of embeddings connecting one to the other. It is well known that two knot diagrams D_1 and D_2 represent ambient isotopic knots if and only if D_1 and D_2 are connected by a finite sequence of the Reidemeister moves, depicted in Figure 1. For more information on these and basic knot theory we refer the reader to the books [Adams 2004; Kauffman 2001; Murasugi 1996; Rolfsen 1976].

On the other hand, two singular link diagrams represent ambient isotopic singular links if their diagrams differ by a finite sequence of the Reidemeister moves together with the extended Reidemeister moves R_4 and R_5 shown in Figure 2; see [Kauffman 1989].

Any knot or link can be assigned an orientation, and there are two possible orientations for a knot and link component. The crossings of an oriented knot will have designated arrows due to the assigned orientation of the knot, and there are two types of crossings, namely positive and negative.



Singular links may also be oriented or unoriented. If a singular link is oriented, then the singular crossings (or four-valent vertices) are crossing-type oriented, which is imposed by the fact that a singular link is an immersion in \mathbb{R}^3 of oriented circles with transversal double points.

In practice, it is tedious to work with Reidemeister moves to determine whether two diagrams represent equivalent knots (or singular links). Instead, one can work with an *invariant* for knots (or singular links), which is a quantity associated to the knot (or singular link) and is independent of the diagram of the knot (or singular link). Equivalently, if K_1 and K_2 are equivalent knots (or singular links), then $Inv(K_1) = Inv(K_2)$ for any invariant Inv . These invariants can be numbers, polynomials, groups, or more complex objects, such as homology theories. In this paper we are concerned with polynomial invariants.

The *Kauffman bracket* [1987] is a polynomial invariant for unoriented knots and links and is defined via a skein relation. A *skein relation* (as in (1-1)) is an identity involving knot diagrams (or singular link diagrams) that are the same except in a small neighborhood where they differ in the way indicated. The Kauffman bracket of a knot diagram K is denoted by $\langle K \rangle$, and is determined by

$$\langle \text{crossing} \rangle = A \langle \text{smooth} \rangle + A^{-1} \langle \text{smooth} \rangle, \tag{1-1}$$

$$\langle \text{circle} \rangle = 1, \quad \langle K \cup \text{circle} \rangle = (-A^2 - A^{-2}) \langle K \rangle. \tag{1-2}$$

It is an enjoyable exercise to show that if two knot diagrams D_1 and D_2 differ by a Reidemeister move $R2$ or $R3$, then $\langle D_1 \rangle = \langle D_2 \rangle$. In other words, the Kauffman bracket is a *regular isotopy invariant* for knots. Note that if an invariant upholds the three Reidemeister moves it is called an *ambient isotopy invariant* for knots.

It is not hard to check that the Kauffman bracket has the following behavior with respect to the Reidemeister move $R1$:

$$\langle \text{twist} \rangle = -A^3 \langle \text{smooth} \rangle \quad \text{and} \quad \langle \text{untwist} \rangle = -A^{-3} \langle \text{smooth} \rangle.$$

By the skein relation defining the Kauffman bracket, every crossing in a knot diagram L is locally replaced with one of the two possible *smoothings*,

$$\left. \begin{array}{c} \text{)} \\ \text{(} \end{array} \right) \left(\text{ or } \begin{array}{c} \text{)} \\ \text{(} \end{array} \right),$$

which will result in a finite number of disjoint circles, called a *state* of K . Note that if K contains n crossings, then there are 2^n states associated with K . The Kauffman bracket polynomial is thus a *state model* polynomial. In this state model, the polynomial $\langle K \rangle \in \mathbb{Z}[A, A^{-1}]$ is given by

$$\langle K \rangle = \sum_{\sigma} A^{\alpha(\sigma) - \beta(\sigma)} (-A^2 - A^{-2})^{\gamma(\sigma) - 1},$$

where the sum is taken over all states σ of the knot diagram K , and where

$$\begin{aligned} \alpha(\sigma) &\text{ is the number of crossings } \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \text{ that have been replaced by } \begin{array}{c} \text{)} \\ \text{(} \end{array}, \\ \beta(\sigma) &\text{ is the number of crossings } \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \text{ that have been replaced by } \begin{array}{c} \text{)} \\ \text{(} \end{array}, \end{aligned}$$

and $\gamma(\sigma)$ is the number of disjoint loops in the state σ . Sometimes, $\alpha(\sigma)$ and $\beta(\sigma)$ are referred to as the numbers of the A -smoothings and A^{-1} -smoothings, respectively.

One can use the Kauffman bracket polynomial to obtain an ambient isotopy invariant for oriented knots by counteracting the behavior of $\langle \cdot \rangle$ with respect to the move $R1$. This is done by defining the Kauffman X polynomial of an oriented knot K given by

$$X(K) := (-A^3)^{-w(K)} \langle K \rangle,$$

where $w(K)$ denotes the *writhe* of the oriented knot diagram K , given by the number of positive crossings minus the number of negative crossings, and where $\langle K \rangle$ is the Kauffman bracket of the unoriented knot diagram obtained from K . Since $w(K)$ and $\langle K \rangle$ are invariant under the moves $R2$ and $R3$, it follows that $X(K)$ is invariant under all three Reidemeister moves. Therefore, the polynomial X is an ambient isotopy invariant for oriented knots.

It is well-known that any polynomial invariant for classical links extends (in various ways) to an invariant of rigid-vertex isotopy for knotted four-valent graphs; see, for example, [Jonish and Millett 1991; Kauffman 1989; 2005; Kauffman and Magarshak 1995; Kauffman and Mishra 2013; Kauffman and Vogel 1992]. (Recall that a singular link can be regarded as a knotted four-valent graph with rigid vertices.) In particular, Kauffman and Vogel [1992] showed that if $I(K)$ is a regular isotopy polynomial invariant for unoriented knots and links, then imposing

the skein relation

$$I\left(\begin{array}{c} \diagup \quad \diagdown \\ \bullet \\ \diagdown \quad \diagup \end{array}\right) = xI\left(\begin{array}{c} \diagup \quad \diagdown \\ \\ \diagdown \quad \diagup \end{array}\right) + xI\left(\begin{array}{c} \diagdown \quad \diagup \\ \\ \diagup \quad \diagdown \end{array}\right) + yI\left(\begin{array}{c} \diagup \quad \diagdown \\ \\ \diagup \quad \diagdown \end{array}\right) + yI\left(\begin{array}{c} \diagdown \quad \diagup \\ \\ \diagdown \quad \diagup \end{array}\right),$$

where x and y are commuting algebraic variables, yields a polynomial invariant, $I(G)$, of rigid-vertex regular isotopy for unoriented knotted graphs G (equivalently, it yields a regular isotopy invariant for unoriented singular links). This method certainly applies to the Kauffman bracket, and we start this paper by borrowing this approach with $x = 1$ and $y = 0$.

We remind the reader that one can also consider a regular isotopy invariant for oriented knots and links and extend it to oriented singular links by applying three local replacements at each singular crossings (that is, at each oriented vertex) and then taking a linear combination of the corresponding replacements. The three replacements are the positive crossing, the negative crossing, and the oriented smoothing at the vertex. For more details on this we refer the reader to [Kauffman 1989; Kauffman and Vogel 1992]. The work in [Kauffman and Magarshak 1995] contains possible applications to molecular biology of invariants of knotted rigid-vertex graphs. More recently, Kauffman and Mishra [2013] introduced a new method for constructing invariants of rigid vertex graph embeddings by using nonlocal combinatorial information that is available at each vertex. In particular, this paper uses the notions of Gauss code and parity for rigid-vertex graphs, and thus it is fundamentally different from the method mentioned earlier.

In this paper we work with a variant of the skein relation above to arrive at a version of the Kauffman bracket for singular links. The main scope of this paper is to show that the resulting polynomial for singular links can be defined in at least two more ways. By providing three approaches to the same polynomial invariant for singular links, we hope that a young researcher reading our paper will find a great deal of information which is educational and interesting, as it reveals beautiful connections between knot theory, combinatorics, abstract algebra, and statistical mechanics.

In Section 2 we give a detailed proof that using $x = 1$ and $y = 0$ in the above skein relation with $I(K) = \langle K \rangle$ yields an invariant for unoriented singular links. We refer to the resulting polynomial as the extended Kauffman bracket. In Section 3 we provide some properties of the extended Kauffman bracket and its associated ambient isotopy invariant for oriented singular links. In Section 4 we define a representation of the singular braid monoid into the Temperley–Lieb algebra, and use it to define a bracket polynomial for singular braids and ultimately recover the extended Kauffman bracket for singular links. Finally, in Section 5 we provide another method for constructing our extended Kauffman bracket; this method relies on a solution to the Yang–Baxter equation. By interpreting singular link diagrams as abstract tensor diagrams, we arrive at a Yang–Baxter state model for the extended Kauffman bracket.

2. An invariant for singular links

In this section, we extend the Kauffman bracket to singular links. For our purpose, we need to associate a skein relation to a singular crossing, and then check that the resulting polynomial is invariant under the extended Reidemeister moves $R4$ and $R5$.

Given a singular link diagram L , we resolve each singular crossing in L using the skein relation

$$\langle \text{Singular Crossing} \rangle = \langle \text{Resolution 1} \rangle + \langle \text{Resolution 2} \rangle. \tag{2-1}$$

This process results in writing $\langle L \rangle$ as a $\mathbb{Z}[A, A^{-1}]$ -linear combination of bracket evaluations of knots and links, which are then evaluated using the rules in (1-1) and (1-2). This yields a Laurent polynomial $\langle L \rangle \in \mathbb{Z}[A, A^{-1}]$.

Note that $\langle L \rangle$ is already invariant under the Reidemeister moves $R2$ and $R3$, since $\langle \cdot \rangle$ is a regular isotopy invariant for knots. Thus, we only need to check that $\langle L \rangle$ is invariant under the moves $R4$ and $R5$. We show this below, where along the way, we use the fact that $\langle \cdot \rangle$ is invariant under the move $R2$ and the behavior of $\langle \cdot \rangle$ with respect to the move $R1$:

$$\langle \text{Singular Crossing} \rangle = \langle \text{Resolution 1} \rangle + \langle \text{Resolution 2} \rangle = \langle \text{Resolution 1} \rangle + \langle \text{Resolution 2} \rangle = \langle \text{Singular Crossing} \rangle.$$

In addition,

$$\begin{aligned} \langle \text{Singular Crossing} \rangle &= \langle \text{Resolution 1} \rangle + \langle \text{Resolution 2} \rangle \\ &= \langle \text{Resolution 1} \rangle + (-A^3)(-A^{-3}) \langle \text{Resolution 2} \rangle = \langle \text{Singular Crossing} \rangle. \end{aligned}$$

It follows that $\langle L \rangle$ is a regular isotopy polynomial invariant for singular links, which we call the *extended Kauffman bracket*. We have proved the statement below.

Theorem 1. *Let L be a singular link diagram and $\langle L \rangle \in \mathbb{Z}[A, A^{-1}]$ be the polynomial given by the following rules:*

$$\begin{aligned} \langle \text{Singular Crossing} \rangle &= \langle \text{Resolution 1} \rangle + \langle \text{Resolution 2} \rangle, \\ \langle \text{Crossing} \rangle &= A \langle \text{Resolution 1} \rangle + A^{-1} \langle \text{Resolution 2} \rangle, \\ \langle \text{Circle} \rangle &= 1, \quad \langle K \cup \text{Circle} \rangle = (-A^2 - A^{-2}) \langle K \rangle. \end{aligned}$$

Then $\langle L \rangle$ is a regular isotopy invariant for L and satisfies

$$\left\langle \begin{array}{c} \text{---} \\ \diagup \quad \diagdown \\ \text{---} \end{array} \right\rangle = -A^3 \left\langle \begin{array}{c} \text{---} \\ \diagdown \quad \diagup \\ \text{---} \end{array} \right\rangle \quad \text{and} \quad \left\langle \begin{array}{c} \text{---} \\ \diagup \quad \diagup \\ \text{---} \end{array} \right\rangle = -A^{-3} \left\langle \begin{array}{c} \text{---} \\ \diagdown \quad \diagdown \\ \text{---} \end{array} \right\rangle.$$

We can define the writhe of an oriented singular link diagram in a similar manner as for the case of oriented knot diagrams. That is, the writhe $w(L)$ of an oriented singular link diagram L is given by the number of positive crossings minus the number of negative crossings. Note that $w(L)$ is independent of the number of singular crossings in L .

Theorem 2. *Let L be an oriented singular link diagram, and let $X(L)$ be the Laurent polynomial defined by*

$$X(L) := (-A^3)^{-w(L)} \langle L \rangle$$

where $\langle L \rangle$ is the extended Kauffman bracket of the unoriented singular link diagram represented by L . Then $X(L)$ is an ambient isotopy invariant for L .

3. Some properties of the extended Kauffman bracket

The goal of this section is to study the behavior of the extended Kauffman bracket polynomial and the polynomial X for singular links with respect to disjoint unions, connected sums, and mirror images of singular links.

For this purpose, we observe first that the extended Kauffman bracket of a singular link can also be defined using a state-sum formula. Let L be a singular link diagram with n classical crossings and m singular crossings. By resolving the classical and singular crossings in L using the first two skein relations in Theorem 1, we write $\langle L \rangle$ as a $\mathbb{Z}[A, A^{-1}]$ -linear combination of bracket evaluations of the states associated with L . Note that L has 2^{n+m} states and that each state is a disjoint union of closed loops. Then

$$\langle L \rangle = \sum_{\sigma} A^{\alpha(\sigma) - \beta(\sigma)} (-A^2 - A^{-2})^{\gamma(\sigma) - 1},$$

where the sum is taken over all states σ associated with the singular link diagram L , where $\gamma(\sigma)$ is the number of disjoint loops in a state σ , and where $\alpha(\sigma)$ and $\beta(\sigma)$ are, respectively, the numbers of A -smoothings and A^{-1} -smoothings in the state σ . (Observe that these smoothings correspond to classical crossings in L .)

Proposition 3. *Let $L_1 \cup L_2$ be the disjoint union of singular link diagrams L_1 and L_2 . Then*

$$\langle L_1 \cup L_2 \rangle = (-A^2 - A^{-2}) \langle L_1 \rangle \langle L_2 \rangle.$$



Figure 3. A pair of disjoint links (left) and their connected sum (right).

Proof. Let $L = L_1 \cup L_2$ and let S be the set of all of the states corresponding to L . We have

$$\langle L \rangle = \langle L_1 \cup L_2 \rangle = \sum_{\sigma \in S} A^{\alpha(\sigma) - \beta(\sigma)} (-A^2 - A^{-2})^{\gamma(\sigma) - 1}.$$

Let S_1 and S_2 represent the set of all of the states associated with L_1 and L_2 , respectively. Observe that the disjoint union of two singular links does not introduce any new crossings and that there is a canonical one-to-one correspondence between $S_1 \times S_2$ and S . For $\sigma_1 \in S_1$, $\sigma_2 \in S_2$, denote by $\sigma \in S$ the state of L which corresponds to (σ_1, σ_2) . Then

$$\alpha(\sigma) = \alpha(\sigma_1) + \alpha(\sigma_2), \quad \beta(\sigma) = \beta(\sigma_1) + \beta(\sigma_2), \quad \gamma(\sigma) = \gamma(\sigma_1) + \gamma(\sigma_2).$$

Therefore,

$$\begin{aligned} \langle L \rangle &= \sum_{\sigma \in S} A^{\alpha(\sigma) - \beta(\sigma)} (-A^2 - A^{-2})^{\gamma(\sigma) - 1} \\ &= \sum_{(\sigma_1, \sigma_2) \in S_1 \times S_2} A^{\alpha(\sigma_1) + \alpha(\sigma_2) - \beta(\sigma_1) - \beta(\sigma_2)} (-A^2 - A^{-2})^{\gamma(\sigma_1) + \gamma(\sigma_2) - 1} \\ &= \sum_{(\sigma_1, \sigma_2) \in S_1 \times S_2} A^{\alpha(\sigma_1) - \beta(\sigma_1)} (-A^2 - A^{-2})^{\gamma(\sigma_1) - 1} A^{\alpha(\sigma_2) - \beta(\sigma_2)} (-A^2 - A^{-2})^{\gamma(\sigma_2) - 1 + 1} \\ &= \langle L_1 \rangle \langle L_2 \rangle (-A^2 - A^{-2}). \quad \square \end{aligned}$$

Corollary 4. Let $L_1 \cup L_2$ be the disjoint union of oriented singular link diagrams L_1 and L_2 . Then,

$$X(L_1 \cup L_2) = (-A^2 - A^{-2})X(L_1)X(L_2).$$

Proof. Note that $w(L_1 \cup L_2) = w(L_1) + w(L_2)$. Combining this and making use of Proposition 3,

$$\begin{aligned} X(L_1 \cup L_2) &= (-A^3)^{-w(L_1 \cup L_2)} \langle L_1 \cup L_2 \rangle \\ &= (-A^3)^{-w(L_1)} \langle L_1 \rangle \cdot (-A^3)^{w(L_2)} \langle L_2 \rangle \cdot (-A^2 - A^{-2}) \\ &= (-A^2 - A^{-2})X(L_1)X(L_2). \quad \square \end{aligned}$$

A singular link diagram L is a *connected sum*, denoted by $L = L_1 \# L_2$, if it is displayed as two disjoint singular link diagrams L_1 and L_2 connected by parallel embedded arcs, up to planar isotopy. Figure 3 shows a connected sum of oriented diagrams.

Proposition 5. *Let L be a singular link diagram with the property that $L = L_1 \# L_2$ for some singular link diagrams L_1 and L_2 . Then the polynomial $\langle L \rangle$ can be computed as*

$$\langle L \rangle = \langle L_1 \rangle \langle L_2 \rangle.$$

Proof. For every state σ of L , there is a pair of states σ_1 and σ_2 of L_1 and L_2 , respectively, such that $\sigma = \sigma_1 \# \sigma_2$. Therefore, $\gamma(\sigma) = \gamma(\sigma_1) + \gamma(\sigma_2) - 1$, while

$$\alpha(\sigma) = \alpha(\sigma_1) + \alpha(\sigma_2) \quad \text{and} \quad \beta(\sigma) = \beta(\sigma_1) + \beta(\sigma_2).$$

Using a similar approach to that in the proof of Proposition 3, we have

$$\begin{aligned} \langle L_1 \# L_2 \rangle &= \sum_{\sigma} A^{\alpha(\sigma) - \beta(\sigma)} (-A^2 - A^{-2})^{\gamma(\sigma) - 1} \\ &= \sum_{\sigma_1} A^{\alpha(\sigma_1) - \beta(\sigma_1)} (-A^2 - A^{-2})^{\gamma(\sigma_1) - 1} \sum_{\sigma_2} A^{\alpha(\sigma_2) - \beta(\sigma_2)} (-A^2 - A^{-2})^{\gamma(\sigma_2) - 1} \\ &= \langle L_1 \rangle \langle L_2 \rangle. \end{aligned} \quad \square$$

Corollary 6. *Let L be an oriented singular link diagram such that $L = L_1 \# L_2$ for some oriented singular link diagrams L_1 and L_2 . Then,*

$$X(L) = X(L_1)X(L_2).$$

Proof. The proof is similar to that of Corollary 4, and thus it is omitted. □

The *mirror image* of a singular link with diagram L is the singular link whose diagram L^* is obtained from L by changing the crossing type for all classical crossings in L . A singular link is *achiral* if it is ambient isotopic to its mirror image and *chiral* otherwise.

Proposition 7. *Let L^* denote the mirror image of a singular link diagram L . Then the extended Kauffman bracket of L^* is obtained from the extended Kauffman bracket of L by interchanging A and A^{-1} . That is,*

$$\langle L^* \rangle(A) = \langle L \rangle(A^{-1}).$$

Proof. According to the state-sum formula defining the extended Kauffman bracket polynomial, it is easy to see that reversing the classical crossings in L replaces an A -smoothing with an A^{-1} -smoothing and vice versa. Hence, the statement follows at once. □

Corollary 8. *If $\langle L \rangle(A) \neq \langle L \rangle(A^{-1})$, then L is a chiral singular link.*

4. A representation of the singular braid monoid

In this section we provide a different approach to the extended Kauffman bracket for singular links, via a representation of the singular braid monoid.

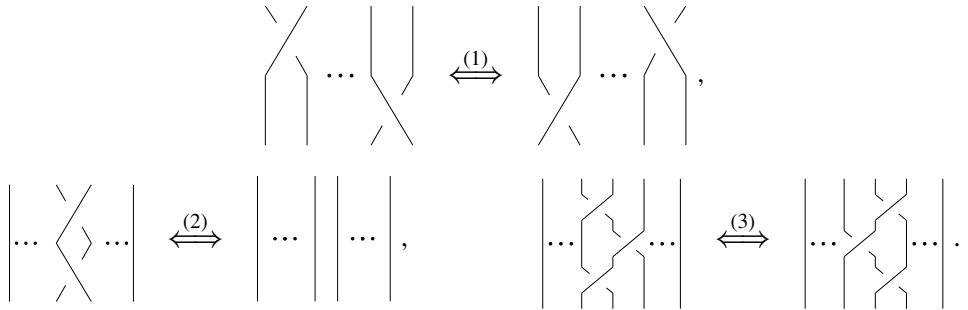
4.1. The singular braid monoid. Let n be a positive integer, $n \geq 2$. Recall that the *singular braid monoid* on n strands, denoted SB_n , is the monoid generated by elements σ_i, σ_i^{-1} , and τ_i , for $1 \leq i \leq n - 1$, where

$$\sigma_i = \left| \begin{array}{c} i \quad i+1 \\ \diagdown \quad \diagup \\ \cdots \quad \cdots \\ \diagup \quad \diagdown \end{array} \right|, \quad \sigma_i^{-1} = \left| \begin{array}{c} i \quad i+1 \\ \diagup \quad \diagdown \\ \cdots \quad \cdots \\ \diagdown \quad \diagup \end{array} \right|, \quad \tau_i = \left| \begin{array}{c} i \quad i+1 \\ \diagdown \quad \diagup \\ \cdots \quad \cdots \\ \diagdown \quad \diagup \end{array} \right|,$$

and satisfying the following relations, under the operation given by vertical concatenation of diagrams:

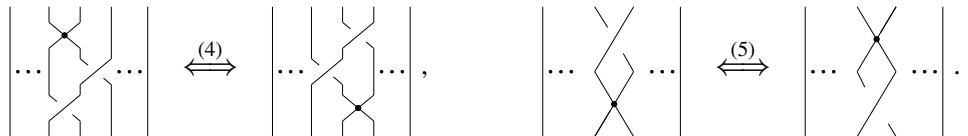
- (1) $g_i h_j = h_j g_i$ for all $g_i, h_i \in \{\sigma_i, \sigma_i^{-1}, \tau_i\}$ and $1 \leq i, j \leq n - 1$ with $|i - j| > 1$,
- (2) $\sigma_i \sigma_i^{-1} = 1_n = \sigma_i^{-1} \sigma_i$ for all $1 \leq i \leq n - 1$,
- (3) $\sigma_i \sigma_j \sigma_i = \sigma_j \sigma_i \sigma_j$ for all $1 \leq i, j \leq n - 1$ with $|i - j| = 1$
- (4) $\tau_i \sigma_j \sigma_i = \sigma_j \sigma_i \tau_j$ for all $1 \leq i, j \leq n - 1$ with $|i - j| = 1$,
- (5) $\sigma_i \tau_i = \tau_i \sigma_i$ for all $1 \leq i \leq n - 1$.

Note that the identity element, denoted 1_n , is represented by n vertical strands with no crossings. The geometric representations of the first three relations is given below (observe that relations (2) and (3) mimic the Reidemeister moves $R2$ and $R3$, respectively):



These three relations (where in (1) we exclude the relations involving the generators τ_i) are exactly the relations in the well-known Artin braid group.

In addition, note that relations (4) and (5) defining the singular braid monoid SB_n mimic, respectively, the moves $R4$ and $R5$ for singular link diagrams:



Due to Joan Birman [1993], we know that every singular link can be expressed as the *closure* of a singular braid, via ambient isotopy. Figure 4 displays the closure $\bar{\beta}$ of a braid β .

There are many different ways to represent a singular link as a closed singular braid. Bernd Gemein [1997] showed that two singular braids have isotopic closures if and only if there exists a finite sequence of singular braid relations and/or extended Markov moves (detailed below) transforming one singular braid into the other.

Let $w \in \text{SB}_n$ be a braid on n strands and let w^* be the natural inclusion of w into SB_{n+1} obtained by adding an $(n+1)$ -st strand to w . Then the following are called the *extended Markov moves*:

- (M1) (a) $\tau_i w \sim w \tau_i$ for all $1 \leq i \leq n - 1$,
- (b) $\sigma_i w \sim w \sigma_i$ for all $1 \leq i \leq n - 1$,
- (M2) $w^* \sigma_n \sim w \sim w^* \sigma_n^{-1}$.

Figure 5 shows isotopic closed braids that differ by an extended Markov move.

Therefore, the works [Birman 1993; Gemein 1997] allow us to relate the theory of singular links with the theory of the singular braid monoid. In particular, we can study the extended Kauffman bracket via the singular braid monoid.

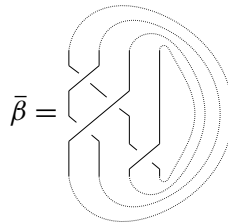


Figure 4. The closure of a braid.

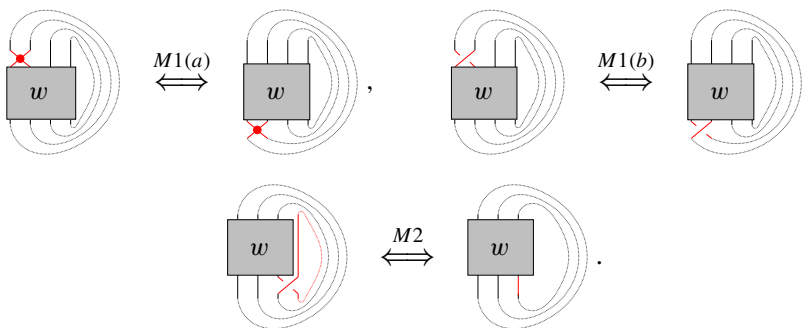


Figure 5. Equivalent singular links under extended Markov moves.

4.2. The Temperley–Lieb algebra. The Temperley–Lieb algebra played a central role in the discovery of the Jones polynomial [1985], and in the subsequent developments relating knot theory, topological quantum field theory, and statistical mechanics [Kauffman 2001]. Originally presented in terms of abstract generators and relations, it was combinatorially described by Kauffman as a planar diagram algebra in terms of his bracket polynomial for unoriented knots.

For each integer $n \geq 2$, the n -strand Temperley–Lieb algebra, denoted TL_n , is the unital, associative algebra over the ring $\mathbb{Z}[A, A^{-1}]$ generated by u_i , for $1 \leq i \leq n - 1$, where

$$u_i = \left| \begin{array}{c} \dots \\ \text{---} \cup \text{---} \\ \dots \\ \dots \\ \text{---} \cap \text{---} \\ \dots \end{array} \right|,$$

along with the identity diagram, denoted 1_n , and subject to the following relations (where multiplication is given by vertical concatenation of diagrams):

- $u_i^2 = (-A^2 - A^{-2})u_i$ for all $1 \leq i \leq n - 1$:

$$\left| \begin{array}{c} \dots \\ \text{---} \cup \text{---} \\ \dots \\ \text{---} \circ \text{---} \\ \dots \\ \text{---} \cap \text{---} \\ \dots \end{array} \right| = (-A^2 - A^{-2}) \left| \begin{array}{c} \dots \\ \text{---} \cup \text{---} \\ \dots \\ \dots \\ \text{---} \cap \text{---} \\ \dots \end{array} \right|,$$

- $u_i u_j u_i = u_i$ for all $1 \leq i, j \leq n - 1$ with $|i - j| = 1$:

$$\left| \begin{array}{c} \dots \\ \text{---} \cup \text{---} \\ \dots \\ \text{---} \cap \text{---} \\ \dots \\ \text{---} \cup \text{---} \\ \dots \\ \text{---} \cap \text{---} \\ \dots \end{array} \right| = \left| \begin{array}{c} \dots \\ \text{---} \cup \text{---} \\ \dots \\ \dots \\ \text{---} \cap \text{---} \\ \dots \end{array} \right|,$$

- $u_i u_j = u_j u_i$ for all $1 \leq i, j \leq n - 1$ with $|i - j| > 1$.

Observe that a generic element in TL_n is a formal $\mathbb{Z}[A, A^{-1}]$ -linear combination of n -strand diagrams formed by multiplications of the generators u_i and the identity 1_n .

Define a trace function $\text{tr} : TL_n \rightarrow \mathbb{Z}[A, A^{-1}]$ given by $\text{tr}(D) = (-A^2 - A^{-2})^{c-1}$, where c is the number of disjoint loops in the diagram \bar{D} obtained by closing the diagram $D \in TL_n$ in the same way that we close a braid or a singular braid. Then extend tr by linearity to all elements of TL_n .

It is easy to see that the function tr satisfies

$$\text{tr}(xy) = \text{tr}(x) \text{tr}(y) \quad \text{for all } x, y \in TL_n. \tag{4-1}$$

4.3. A representation of \mathbf{SB}_n . We observe that for any given homomorphism $\rho : \mathbf{SB}_n \rightarrow \mathbf{TL}_n$, we can compose it with the trace function tr to obtain, for any singular braid element $w \in \mathbf{SB}_n$, a polynomial $(\text{tr} \circ \rho)(w) \in \mathbb{Z}[A, A^{-1}]$.

Inspired by the skein relations defining the extended Kauffman bracket for singular links, we define a homomorphism $\rho : \mathbf{SB}_n \rightarrow \mathbf{TL}_n$ as follows:

$$\begin{aligned} \tau_i &\xrightarrow{\rho} u_i + 1_n, \\ \sigma_i &\xrightarrow{\rho} A^{-1}u_i + A1_n, \\ \sigma_i^{-1} &\xrightarrow{\rho} Au_i + A^{-1}1_n. \end{aligned}$$

We can think of ρ as a function that resolves the crossings of the singular braid, since each σ_i , σ_i^{-1} , and τ_i represents a crossing of the strands in the singular braid.

Theorem 9. *The map ρ is a representation of the singular braid monoid \mathbf{SB}_n into the Temperley–Lieb algebra \mathbf{TL}_n . That is, ρ preserves the singular braid monoid relations.*

Proof. First, observe that ρ preserves the commuting relations in \mathbf{SB}_n , since the generators for the algebra \mathbf{TL}_n satisfy similar commuting relations. Note also that it must be the case that ρ preserves the relations (2)–(5) in \mathbf{SB}_n , since the extended Kauffman bracket is invariant under the Reidemeister moves $R2$ and $R3$, as well as under the moves $R4$ and $R5$. However, we will check two of the singular braid monoid relations and leave the other relations as an exercise.

We start off by verifying that $\rho(\tau_i\sigma_j\sigma_i) = \rho(\sigma_j\sigma_i\tau_j)$. First, observe that

$$\begin{aligned} \rho(\tau_i\sigma_j\sigma_i) &= \rho(\tau_i)\rho(\sigma_j)\rho(\sigma_i) \\ &= (u_i + 1_n)(A^{-1}u_j + A1_n)(A^{-1}u_i + A1_n), \\ \rho(\sigma_j\sigma_i\tau_j) &= \rho(\sigma_j)\rho(\sigma_i)\rho(\tau_j) \\ &= (A^{-1}u_j + A1_n)(A^{-1}u_i + A1_n)(u_j + 1_n). \end{aligned}$$

Employing the relations in \mathbf{TL}_n , we have

$$\begin{aligned} &(u_i + 1_n)(A^{-1}u_j + A1_n)(A^{-1}u_i + A1_n) \\ &= A^2u_i + u_i^2 + A^{-2}u_iu_ju_i + u_iu_j + A^{-2}u_ju_i + A^21_n + u_i + u_j \\ &= A^2u_i + (-A^2 - A^{-2})u_i + A^{-2}u_i + u_iu_j + A^{-2}u_ju_i + A^21_n + u_i + u_j \\ &= u_iu_j + A^{-2}u_ju_i + A^21_n + u_i + u_j \\ &= A^2u_j + (-A^2 - A^{-2})u_j + A^{-2}u_j + u_iu_j + A^{-2}u_ju_i + A^21_n + u_i + u_j \\ &= A^2u_j + u_j^2 + A^{-2}u_ju_iu_j + u_iu_j + A^{-2}u_ju_i + A^21_n + u_i + u_j \\ &= (A^{-1}u_j + A1_n)(A^{-1}u_i + A1_n)(u_j + 1_n). \end{aligned}$$

It follows that the fourth relation defining SB_n is preserved by the map ρ . Next we show that $\rho(\tau_i\sigma_i) = \rho(\sigma_i\tau_i)$. Using basic computations, we obtain

$$\begin{aligned}\rho(\tau_i\sigma_i) &= \rho(\tau_i)\rho(\sigma_i) = (u_i + 1_n)(A^{-1}u_i + A1_n) \\ &= A^{-1}u_i^2 + A^{-1}u_i + Au_i + A1_n \\ &= (A^{-1}u_i + A1_n)(-A^{-3})(u_i + 1_n) \\ &= \rho(\sigma_i)\rho(\tau_i) = \rho(\sigma_i\tau_i).\end{aligned}$$

This shows that ρ also preserves the fifth relation defining SB_n . \square

Remark 10. For any $a, b \in \mathbb{Z}[A, A^{-1}]$, the homomorphism $f : \text{SB}_n \rightarrow \text{TL}_n$ given by

$$\tau_i \xrightarrow{f} au_i + b1_n \quad \text{and} \quad \sigma_i^{\pm 1} \xrightarrow{f} A^{\mp 1}u_i + A^{\pm 1}1_n$$

also defines a representation of the singular braid monoid SB_n into the Temperley–Lieb algebra TL_n . The proof that f preserves the singular braid monoid relations follows verbatim as that for the map ρ .

4.4. The bracket polynomial of a singular braid. In this section, we show how to recover the extended Kauffman bracket of singular links by making use of the map ρ and the trace function tr .

Let $\beta \in \text{SB}_n$ be a singular braid on n strands and denote by $\text{wr}(\beta)$ the writhe of β , defined as the sum of the number of generators of type σ_i minus the sum of the generators of type σ_i^{-1} in the expression of β .

Define the function $\langle \cdot \rangle : \text{SB}_n \rightarrow \mathbb{Z}[A, A^{-1}]$, given by the formula

$$\langle \beta \rangle = (-A^3)^{-\text{wr}(\beta)} (\text{tr} \circ \rho)(\beta).$$

We call $\langle \beta \rangle$ the bracket polynomial of the singular braid β .

Proposition 11. *The bracket polynomial of a singular braid is well-defined on singular braids, and is invariant under the extended Markov moves. Moreover, if L is a singular link diagram in braid form such that $L = \bar{\beta}$ for some $\beta \in \text{SB}_n$, then*

$$\langle \beta \rangle = (-A^3)^{-\text{wr}(\beta)} \langle \bar{\beta} \rangle = (-A^3)^{-\text{wr}(\beta)} \langle L \rangle.$$

Proof. Since ρ is a representation of SB_n and the writhe of the singular braid is invariant under the relations in SB_n , it follows that the bracket polynomial of a singular braid is well-defined on singular braids. The trace function tr satisfies (4-1), and thus

$$(\text{tr} \circ \rho)(\tau_i w) = \text{tr}(\rho(\tau_i)\rho(w)) = \text{tr}(\rho(w)\rho(\tau_i)) = (\text{tr} \circ \rho)(w\tau_i),$$

and similarly,

$$(\text{tr} \circ \rho)(\sigma_i w) = (\text{tr} \circ \rho)(w\sigma_i)$$

for all $\tau_i, \sigma_i, w \in \text{SB}_n$. Thus $\langle \cdot \rangle$ is invariant under the extended Markov moves of type (M1). Moreover, the coefficient $(-A^3)^{-\text{wr}(\beta)}$ in the expression of $\langle \cdot \rangle$ cancels the effect of a Markov move of type (M2):

$$\langle w^* \sigma_n \rangle = (-A^3)^{-\text{wr}(\beta)-1} (\text{tr} \circ \rho)(w^* \sigma_n) = (-A^3)^{-\text{wr}(\beta)} (\text{tr} \circ \rho)(w) = \langle w \rangle.$$

Finally, due to [Birman 1993; Gemein 1997] and the definitions for the maps ρ and tr , the second part of the statement follows immediately. \square

5. The Yang–Baxter equation and the extended Kauffman bracket

We will show now how to arrive at the extended Kauffman bracket by interpreting singular link diagrams as *abstract tensor diagrams* and employing a solution to the Yang–Baxter equation.

5.1. A Yang–Baxter model for the extended Kauffman bracket. Our approach here is an extension from classical knots to singular links of the Yang–Baxter state model for the Kauffman bracket, as introduced in [Kauffman 2001].

A singular link diagram D can be decomposed with respect to a height function into minima (creations), maxima (annihilations) and crossings (interactions), as illustrated in Figure 6. That is, the diagram D is constructed from interconnected maxima, minima, and crossings (there might be some curves with no critical points vis-a-vis the height function), and we want to associate to them square matrices with entries in the ring $\mathbb{Z}[A, A^{-1}]$. We start by labeling the edges of the diagram D with *spins* from the index set $I = \{1, 2\}$.

We will denote the following portions of the link diagram as follows:

$$\begin{aligned} M_{a,b} &\longleftrightarrow a \text{ (arc over) } b, & M^{a,b} &\longleftrightarrow a \text{ (arc under) } b, \\ R_{c,d}^{a,b} &\longleftrightarrow \begin{array}{c} a \quad b \\ \diagdown \quad \diagup \\ c \quad d \end{array}, & \bar{R}_{c,d}^{a,b} &\longleftrightarrow \begin{array}{c} a \quad b \\ \diagup \quad \diagdown \\ c \quad d \end{array}, \\ Q_{c,d}^{a,b} &\longleftrightarrow \begin{array}{c} a \quad b \\ \diagup \quad \diagdown \\ c \quad d \end{array}, & \delta_b^a &\longleftrightarrow \begin{array}{c} a \\ | \\ b \end{array}, \end{aligned}$$

where $a, b, c, d \in I$ and

$$\delta_b^a = \begin{cases} 1, & \text{if } a = b, \\ 0, & \text{if } a \neq b. \end{cases}$$

Using these conventions, we wish to associate to any singular link diagram D a polynomial $\tau(D) \in \mathbb{Z}[A, A^{-1}]$ so that $\tau(D)$ recovers the extended Kauffman bracket $\langle D \rangle$. The expression $\tau(D)$ is obtained by taking the sum over all internal labels (spins on the arcs of the diagram D) of the products of symbols representing maxima, minima, and crossings (classical and singular).

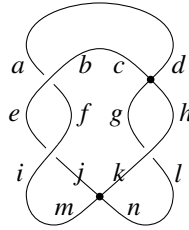


Figure 6. An abstract tensor singular link diagram.

For example, for the diagram D in Figure 6, $\tau(D)$ is given by the following sum of products of abstract tensor symbols:

$$\tau(D) = \sum_{a,b,\dots,n \in I} M_{a,d} M_{b,c} R_{e,f}^{a,b} R_{i,j}^{e,f} M^{i,m} Q_{m,n}^{j,k} M^{n,l} R_{k,l}^{g,h} Q_{g,h}^{c,d},$$

where the sum is over all possible choices of indices (spins from I) in the expression. Note that the order of the factors in a product of abstract tensors does not matter, since the abstract tensors are elements of the commutative ring $\mathbb{Z}[A, A^{-1}]$.

We will use the following notational conventions:

$$X = (X)_{c,d}^{a,b} = \begin{bmatrix} X_{1,1}^{1,1} & X_{1,2}^{1,1} & X_{2,1}^{1,1} & X_{2,2}^{1,1} \\ X_{1,1}^{1,2} & X_{1,2}^{1,2} & X_{2,1}^{1,2} & X_{2,2}^{1,2} \\ X_{1,1}^{2,1} & X_{1,2}^{2,1} & X_{2,1}^{2,1} & X_{2,2}^{2,1} \\ X_{1,1}^{2,2} & X_{1,2}^{2,2} & X_{2,1}^{2,2} & X_{2,2}^{2,2} \end{bmatrix}$$

and

$$(B)_{a,b} = (B)^{a,b} = (B)_b^a = \begin{bmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{bmatrix}.$$

Observe that

$$\sum_{c,d \in I} X_{c,d}^{a,b} Y_{e,f}^{c,d} = (XY)_{e,f}^{a,b} \quad \text{for all } a, b, e, f \in I.$$

This can be easily seen by rewriting $X_{c,d}^{a,b}$ as X_j^i , where $i = b + 2(a - 1)$ and $j = d + 2(c - 1)$, since

$$\sum_{j=1}^4 X_j^i Y_k^j = (XY)_k^i.$$

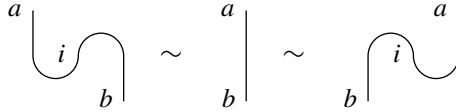
To arrive at the bracket polynomial, the matrices corresponding to maxima and minima need to satisfy

$$\sum_{a,b \in I} M_{a,b} M^{a,b} \longleftrightarrow \bigcirc \longleftrightarrow -A^2 - A^{-2}.$$

By imposing $M_{a,b} = M^{a,b}$ for $a, b \in I$, the above equality becomes

$$\sum_{a,b \in I} (M_{a,b})^2 = -A^2 - A^{-2} = \sum_{a,b \in I} (M^{a,b})^2. \tag{5-1}$$

Since we want $\tau(D)$ to be a topological invariant, pairs of maxima and minima should cancel as shown:



Therefore, we need that

$$\sum_{i \in I} M^{a,i} M_{i,b} = \delta_b^a = \sum_{i \in I} M_{b,i} M^{i,a}$$

or, equivalently,

$$\sum_{i \in I} M_{a,i} M_{i,b} = \delta_b^a = \sum_{i \in I} M_{b,i} M_{i,a}. \tag{5-2}$$

It follows that the matrix $M = (M_{a,b})$ should be its own inverse. The following matrix satisfies (5-1) and (5-2):

$$M = \begin{bmatrix} 0 & iA \\ -iA^{-1} & 0 \end{bmatrix}, \quad \text{where } i^2 = -1.$$

We wish $\tau(D)$ to satisfy the Kauffman bracket skein relation

$$\tau \left(\begin{array}{c} \diagdown \quad \diagup \\ \diagup \quad \diagdown \end{array} \right) = A \tau \left(\begin{array}{c} \diagdown \quad \diagdown \\ \diagup \quad \diagup \end{array} \right) + A^{-1} \tau \left(\begin{array}{c} \diagup \quad \diagup \\ \diagdown \quad \diagdown \end{array} \right)$$

and thus the R -matrix should satisfy

$$R_{c,d}^{a,b} \longleftrightarrow \begin{array}{c} a \quad b \\ \diagdown \quad \diagup \\ c \quad d \end{array} = A \begin{array}{c} a \\ \diagdown \quad \diagdown \\ c \end{array} + A^{-1} \begin{array}{c} a \quad b \\ \diagup \quad \diagup \\ c \quad d \end{array}.$$

Therefore,

$$R_{c,d}^{a,b} = A \delta_c^a \delta_d^b + A^{-1} M^{a,b} M_{c,d} \quad \text{for all } a, b, c, d \in I.$$

Note that the matrix $U = (U_{c,d}^{a,b}) := (M^{a,b} M_{c,d})$, where

$$U_{c,d}^{a,b} = \begin{array}{c} a \quad b \\ \diagup \quad \diagup \\ c \quad d \end{array}$$

has the following expression:

$$U = \begin{bmatrix} M^{1,1}M_{1,1} & M^{1,1}M_{1,2} & M^{1,1}M_{2,1} & M^{1,1}M_{2,2} \\ M^{1,2}M_{1,1} & M^{1,2}M_{1,2} & M^{1,2}M_{2,1} & M^{1,2}M_{2,2} \\ M^{2,1}M_{1,1} & M^{2,1}M_{1,2} & M^{2,1}M_{2,1} & M^{2,1}M_{2,2} \\ M^{2,2}M_{1,1} & M^{2,2}M_{1,2} & M^{2,2}M_{2,1} & M^{2,2}M_{2,2} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -A^2 & 1 & 0 \\ 0 & 1 & -A^{-2} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Moreover, observe that

$$(\delta_c^a \delta_d^b) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = (\delta_c^a) \otimes (\delta_d^b),$$

where \otimes represents the Kronecker product of matrices.

Furthermore, the \bar{R} -matrix should satisfy

$$\bar{R}_{c,d}^{a,b} \longleftrightarrow \begin{array}{c} a \quad b \\ \diagdown \quad / \\ c \quad d \end{array} = A \begin{array}{c} a \quad b \\ \frown \quad \smile \\ c \quad d \end{array} + A^{-1} \begin{array}{c} a \\ | \\ c \end{array} \begin{array}{c} b \\ | \\ d \end{array}$$

and thus

$$\bar{R}_{c,d}^{a,b} = AM^{a,b}M_{c,d} + A^{-1}\delta_c^a\delta_d^b \quad \text{for all } a, b, c, d \in I.$$

We arrive at the following matrices associated with classical crossings:

$$R = \begin{bmatrix} A & 0 & 0 & 0 \\ 0 & 0 & A^{-1} & 0 \\ 0 & A^{-1} & A-A^{-3} & 0 \\ 0 & 0 & 0 & A \end{bmatrix} \quad \text{and} \quad \bar{R} = \begin{bmatrix} A^{-1} & 0 & 0 & 0 \\ 0 & A^{-1}-A^3 & A & 0 \\ 0 & A & 0 & 0 \\ 0 & 0 & 0 & A^{-1} \end{bmatrix}.$$

Finally, we wish $\tau(D)$ to also satisfy

$$\tau\left(\begin{array}{c} a \quad b \\ \diagdown \quad / \\ c \quad d \end{array}\right) = \tau\left(\begin{array}{c} a \\ | \\ c \end{array}\right) \left(\begin{array}{c} b \\ | \\ d \end{array}\right) + \tau\left(\begin{array}{c} a \quad b \\ \frown \quad \smile \\ c \quad d \end{array}\right)$$

which forces the matrix Q associated with a singular crossing to be given by

$$Q_{c,d}^{a,b} = \delta_c^a \delta_d^b + M^{a,b}M_{c,d} \quad \text{for all } a, b, c, d \in I.$$

Equivalently,

$$Q = (Q_{c,d}^{a,b}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-A^2 & 1 & 0 \\ 0 & 1 & 1-A^{-2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Now that we have defined $\tau(D)$ for a given singular link diagram D , we need to make sure that it is a regular isotopy invariant for D . That is, we need to verify

that if D_1 and D_2 are singular link diagrams that differ by a Reidemeister move $R2$ or $R3$, or by an extended Reidemeister move $R4$ or $R5$, then $\tau(D_1) = \tau(D_2)$.

An easy check shows that matrices R and \bar{R} are inverses of each other, since $R\bar{R} = I_{4 \times 4} = \bar{R}R$. Equivalently,

$$\sum_{i,j \in I} R_{i,j}^{a,b} \bar{R}_{c,d}^{i,j} \longleftrightarrow \left(\begin{array}{c} \text{---} \\ \text{---} \end{array} \right) \sim \left(\begin{array}{c} \text{---} \\ \text{---} \end{array} \right) \longleftrightarrow \delta_c^a \delta_d^b,$$

$$\sum_{i,j \in I} \bar{R}_{i,j}^{a,b} R_{c,d}^{i,j} \longleftrightarrow \left(\begin{array}{c} \text{---} \\ \text{---} \end{array} \right) \sim \left(\begin{array}{c} \text{---} \\ \text{---} \end{array} \right) \longleftrightarrow \delta_c^a \delta_d^b.$$

Hence, $\tau(D)$ is invariant under the Reidemeister move $R2$. Moreover, we have that

$$\sum_{i,j,k \in I} R_{i,j}^{a,b} R_{k,f}^{j,c} R_{d,e}^{i,k} \longleftrightarrow \left(\begin{array}{c} \text{---} \\ \text{---} \end{array} \right) \sim \left(\begin{array}{c} \text{---} \\ \text{---} \end{array} \right) \longleftrightarrow \sum_{i,j,k \in I} R_{i,j}^{b,c} R_{d,k}^{a,i} R_{e,f}^{k,j}.$$

The latter relation is the *Yang–Baxter equation* (YBE):

$$\sum_{i,j,k \in I} R_{i,j}^{a,b} R_{k,f}^{j,c} R_{d,e}^{i,k} = \sum_{i,j,k \in I} R_{i,j}^{b,c} R_{d,k}^{a,i} R_{e,f}^{k,j},$$

which can be rewritten as

$$(R \otimes I)(I \otimes R)(R \otimes I) = (I \otimes R)(R \otimes I)(I \otimes R).$$

That is, the R -matrix as defined above is a solution of the YBE. Similarly, one can easily verify that the matrix \bar{R} is a solution of the YBE. It follows that $\tau(D)$ is invariant under the Reidemeister move $R3$.

Furthermore, it is not hard to check that the following holds:

$$\sum_{i,j,k \in I} Q_{i,j}^{a,b} R_{k,f}^{j,c} R_{d,e}^{i,k} \longleftrightarrow \left(\begin{array}{c} \text{---} \\ \text{---} \end{array} \right) \sim \left(\begin{array}{c} \text{---} \\ \text{---} \end{array} \right) \longleftrightarrow \sum_{i,j,k \in I} R_{i,j}^{b,c} R_{d,k}^{a,i} Q_{e,f}^{k,j},$$

or, equivalently,

$$(Q \otimes I)(I \otimes R)(R \otimes I) = (I \otimes R)(R \otimes I)(I \otimes Q).$$

A similar relation holds for R being replaced by \bar{R} . Hence, $\tau(D)$ is invariant under the extended Reidemeister move $R4$.

Finally, observe that $RQ = QR$ and $\bar{R}Q = Q\bar{R}$, or equivalently,

$$RQ\bar{R} = Q \quad \text{and} \quad \bar{R}QR = Q.$$

Therefore, $\tau(D)$ is invariant under the extended Reidemeister move $R5$.

According to the above discussion, we have proved the following statement.

Theorem 12. *The polynomial $\tau(D) \in \mathbb{Z}[A, A^{-1}]$ is an invariant of regular isotopy for singular links.*

Remark 13. We note that $\tau(D)$ is the unnormalized extended Kauffman bracket. That is,

$$\tau(D) = (-A^2 - A^{-2})\langle D \rangle,$$

where $\langle D \rangle$ is the extended Kauffman bracket introduced in Section 2.

5.2. Yet another representation of SB_n . We can use the matrices R, \bar{R} , and Q to define a representation of the singular braid monoid SB_n into a matrix algebra over the ring $\mathbb{Z}[A, A^{-1}]$. Observe first that we can regard a generator for SB_n as an abstract tensor diagram. For example,

$$\sigma_i = \left| \begin{array}{c} \cdots \quad \diagdown \quad \diagup \quad \cdots \\ \cdots \quad \diagup \quad \diagdown \quad \cdots \end{array} \right| \longleftrightarrow \delta_{b_1}^{a_1} \cdots R_{b_i, b_{i+1}}^{a_i, a_{i+1}} \cdots \delta_{b_n}^{a_n} \in M_{2^n \times 2^n}(\mathbb{Z}[A, A^{-1}]).$$

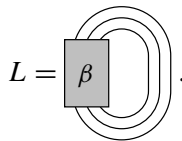
Inspired by this, we define a homomorphism $\Delta : \text{SB}_n \rightarrow M_{2^n \times 2^n}(\mathbb{Z}[A, A^{-1}])$ given by

$$\begin{aligned} \sigma_i &\mapsto I^{\otimes(i-1)} \otimes R \otimes I^{\otimes(n-i-1)}, \\ \sigma_i^{-1} &\mapsto I^{\otimes(i-1)} \otimes \bar{R} \otimes I^{\otimes(n-i-1)}, \\ \tau_i &\mapsto I^{\otimes(i-1)} \otimes Q \otimes I^{\otimes(n-i-1)}. \end{aligned}$$

Since the polynomial $\tau(D)$ is a regular isotopy invariant for singular links, it follows that the map Δ preserves the singular braid monoid relations. Therefore, the following statement holds.

Proposition 14. *The mapping Δ is a representation of the singular braid monoid SB_n into the matrix algebra $M_{2^n \times 2^n}(\mathbb{Z}[A, A^{-1}])$.*

The mapping Δ provides yet another method for obtaining the extended bracket polynomial of a singular link. Let L be a singular link diagram in braid form and let $\beta \in \text{SB}_n$ be the singular braid whose closure is L . That is, $L = \bar{\beta}$:



When closing a braid, each braid strand contributes a diagram and an associated matrix of the form

$$\eta_b^a = \left(\begin{array}{c} a \\ \text{---} \\ b \end{array} \right) \longleftrightarrow \sum_{c \in I} M_{a,c} M^{b,c}, \quad \text{where } a, b \in I.$$

The matrix $\eta = (\eta_b^a)$ is

$$\eta = \begin{bmatrix} M_{1,1}M^{1,1} + M_{1,2}M^{1,2} & M_{1,1}M^{2,1} + M_{1,2}M^{2,2} \\ M_{2,1}M^{1,1} + M_{2,2}M^{1,2} & M_{2,1}M^{2,1} + M_{2,2}M^{2,2} \end{bmatrix} = \begin{bmatrix} -A^2 & 0 \\ 0 & -A^{-2} \end{bmatrix}.$$

Observe that $\text{Trace}(\eta)$, the trace of the matrix η , is $-A^2 - A^{-2}$. Moreover,

$$\tau(L) = \text{Trace}(\eta^{\otimes n} \Delta(\beta)),$$

whenever $\beta \in \text{SB}_n$ and $\bar{\beta} = L$.

Acknowledgements

This research was completed during the 2013 Fresno State Mathematics REU Program, supported by NSF grant #DMS-1156273. The authors would also like to thank the referee for a careful reading of the paper and valuable comments and suggestions.

References

- [Adams 2004] C. C. Adams, *The knot book: an elementary introduction to the mathematical theory of knots*, 2nd ed., American Mathematical Society, Providence, RI, 2004. MR Zbl
- [Birman 1993] J. S. Birman, “New points of view in knot theory”, *Bull. Amer. Math. Soc. (N.S.)* **28:2** (1993), 253–287. MR Zbl
- [Gemein 1997] B. Gemein, “Singular braids and Markov’s theorem”, *J. Knot Theory Ramifications* **6:4** (1997), 441–454. MR Zbl
- [Jones 1985] V. F. R. Jones, “A polynomial invariant for knots via von Neumann algebras”, *Bull. Amer. Math. Soc. (N.S.)* **12:1** (1985), 103–111. MR Zbl
- [Jonish and Millett 1991] D. Jonish and K. C. Millett, “Isotopy invariants of graphs”, *Trans. Amer. Math. Soc.* **327:2** (1991), 655–702. MR Zbl
- [Kauffman 1987] L. H. Kauffman, “State models and the Jones polynomial”, *Topology* **26:3** (1987), 395–407. MR Zbl
- [Kauffman 1989] L. H. Kauffman, “Invariants of graphs in three-space”, *Trans. Amer. Math. Soc.* **311:2** (1989), 697–710. MR Zbl
- [Kauffman 2001] L. H. Kauffman, *Knots and physics*, 3rd ed., Series on Knots and Everything **1**, World Scientific, River Edge, NJ, 2001. MR Zbl
- [Kauffman 2005] L. H. Kauffman, “Knot diagrammatics”, pp. 233–318 in *Handbook of knot theory*, edited by W. Menasco and M. Thistlethwaite, Elsevier, Amsterdam, 2005. MR Zbl
- [Kauffman and Magarshak 1995] L. H. Kauffman and Y. B. Magarshak, “Vassiliev knot invariants and the structure of RNA folding”, pp. 343–394 in *Knots and applications*, edited by L. H. Kauffman, Series on Knots and Everything **6**, World Scientific, River Edge, NJ, 1995. MR Zbl
- [Kauffman and Mishra 2013] L. H. Kauffman and R. Mishra, “Nodal parity invariants of knotted rigid vertex graphs”, *J. Knot Theory Ramifications* **22:4** (2013), Article ID #1340002. MR Zbl
- [Kauffman and Vogel 1992] L. H. Kauffman and P. Vogel, “Link polynomials and a graphical calculus”, *J. Knot Theory Ramifications* **1:1** (1992), 59–104. MR Zbl

[Murasugi 1996] K. Murasugi, *Knot theory and its applications*, Birkhäuser, Boston, MA, 1996. MR Zbl

[Rolfsen 1976] D. Rolfsen, *Knots and links*, Mathematics Lecture Series 7, Publish or Perish, Berkeley, CA, 1976. Revised edition published by the American Mathematical Society, Providence, RI, 2003. MR Zbl

Received: 2014-06-29 Revised: 2015-01-28 Accepted: 2015-08-17

ccaprau@csufresno.edu *Department of Mathematics, California State University,
Fresno, 5245 N. Backer Avenue M/S PB108,
Fresno, CA 93740-8001, United States*

alexchi1114@gmail.com *Department of Mathematics, SUNY Geneseo,
Geneseo, NY 14454, United States*

pyc3@ricealumni.net *Department of Mathematics, Rice University,
Houston, TX 77005, United States*

Symplectic embeddings of four-dimensional ellipsoids into polydiscs

Madeleine Burkhart, Prieră Panescu and Max Timmons

(Communicated by Bjorn Poonen)

McDuff and Schlenk recently determined exactly when a four-dimensional symplectic ellipsoid symplectically embeds into a symplectic ball. Similarly, Frenkel and Müller recently determined exactly when a symplectic ellipsoid symplectically embeds into a symplectic cube. Symplectic embeddings of more complicated sets, however, remain mostly unexplored. We study when a symplectic ellipsoid $E(a, b)$ symplectically embeds into a polydisc $P(c, d)$. We prove that there exists a constant C depending only on d/c (here, d is assumed greater than c) such that if b/a is greater than C , then the only obstruction to symplectically embedding $E(a, b)$ into $P(c, d)$ is the volume obstruction. We also conjecture exactly when an ellipsoid embeds into a scaling of $P(1, b)$ for $b \geq 6$, and conjecture about the set of (a, b) such that the only obstruction to embedding $E(1, a)$ into a scaling of $P(1, b)$ is the volume. Finally, we verify our conjecture for $b = \frac{13}{2}$.

1. Introduction	219
2. Proof of Theorem 1.1	222
3. Proof of Theorem 1.2, Part I	224
4. Proof of Theorem 1.2, Part II	228
5. Proof of Theorem 1.2, Part III	236
6. Conjectures	238
Acknowledgments	241
References	242

1. Introduction

Statement of results. Let (X_0, ω_0) and (X_1, ω_1) be symplectic manifolds. A *symplectic embedding* of (X_0, ω_0) into (X_1, ω_1) is a smooth embedding φ such that

MSC2010: 53DXX.

Keywords: symplectic geometry.

$\varphi^*(\omega_1) = \omega_0$. It is interesting to ask when one symplectic manifold embeds into another. For example, define the (open) four-dimensional symplectic *ellipsoid*

$$E(a, b) = \left\{ (z_1, z_2) \in \mathbb{C}^2 : \frac{\pi|z_1|^2}{a} + \frac{\pi|z_2|^2}{b} < 1 \right\}, \quad (1-1)$$

and define the (open) *symplectic ball* $B(a) := E(a, a)$. These inherit symplectic forms by restricting the standard form $\omega = \sum_{k=1}^2 dx_k dy_k$ on $\mathbb{R}^4 = \mathbb{C}^2$. McDuff and Schlenk [2012] determined exactly when a four-dimensional symplectic ellipsoid $E(a, b)$ embeds symplectically into a symplectic ball, and found that if b/a is small, then the answer involves an “infinite staircase” determined by the Fibonacci numbers with odd index, while if b/a is large then all obstructions vanish except for the volume obstruction.

To give another example, define the (open) four-dimensional *polydisc*

$$P(a, b) = \{ (z_1, z_2) \in \mathbb{C}^2 : \pi|z_1|^2 < a, \pi|z_2|^2 < b \}, \quad (1-2)$$

where $a, b \geq 1$ are real numbers and the symplectic form is again given by restricting the standard symplectic form on \mathbb{R}^4 . Frenkel and Müller [2012] determined exactly when a four-dimensional symplectic ellipsoid symplectically embeds into a *cube* $C(a) := P(a, a)$ and found that part of the expression involves the Pell numbers. Cristofaro-Gardiner and Kleinman [2013] studied embeddings of four-dimensional ellipsoids into scalings of $E(1, \frac{3}{2})$ and also found that part of the answer involves an infinite staircase determined by a recursive sequence.

Here we study symplectic embeddings of an open four-dimensional symplectic ellipsoid $E(a, b)$ into an open four-dimensional symplectic polydisc $P(c, d)$. By scaling, we can encode this embedding question as the function

$$d(a, b) := \inf \{ \lambda : E(1, a) \xrightarrow{c} P(\lambda, b\lambda) \}, \quad (1-3)$$

where a and b are real numbers that are both greater than or equal to 1.

The function $d(a, b)$ always has a lower bound, $\sqrt{a/(2b)}$, the volume obstruction. Our first theorem states that for fixed b , if a is sufficiently large then this lower bound is sharp, i.e., all embedding obstructions vanish aside from the volume obstruction:

Theorem 1.1. *If $a \geq 9(b+1)^2/(2b)$, then $d(a, b) = \sqrt{a/(2b)}$.*

This is an analogue of a result of Buse and Hind [2013] concerning symplectic embeddings of one symplectic ellipsoid into another.

From the previously mentioned work of McDuff and Schlenk, Frenkel and Müller, and Cristofaro-Gardiner and Kleinman, one expects that if a is small then the function $d(a, b)$ should be more rich. Our results suggest that this is indeed the case. For example, we completely determine the graph of $d(a, \frac{13}{2})$ (see Figure 1).

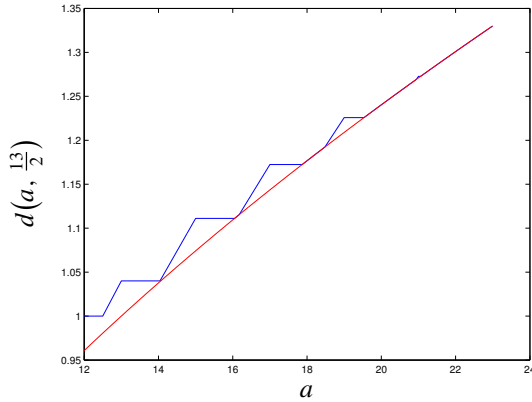


Figure 1. The graph of $d(a, \frac{13}{2})$. The red line represents the volume obstruction.

Theorem 1.2. For $b = \frac{13}{2}$, we have $d(a, b) \geq \sqrt{a/13}$ and is equal to this lower bound for all a except in the following cases:

- (i) $d(a, \frac{13}{2}) = 1$ for all $a \in [1, \frac{25}{2}]$.
- (ii) For $k \in \mathbb{Z}$, with $0 \leq k \leq 4$,

$$d(a, b) = \begin{cases} \frac{2a}{25+2k} & \text{if } a \in [\alpha_k, 13+2k], \\ \frac{26+4k}{25+2k} & \text{if } a \in [13+2k, \beta_k], \end{cases}$$

where

$$\begin{aligned} \alpha_0 &= \frac{25}{2}, & \alpha_1 &= \frac{351}{25}, & \alpha_2 &= \frac{841}{52}, & \alpha_3 &= \frac{961}{52}, & \alpha_4 &= \frac{1089}{52}, \\ \beta_0 &= \frac{351}{25}, & \beta_1 &= \frac{1300}{81}, & \beta_2 &= \frac{15028}{841}, & \beta_3 &= \frac{18772}{961}, & \beta_4 &= \frac{2548}{121}. \end{aligned}$$

Interestingly, the graph of $d(a, \frac{13}{2})$ has only finitely many nonsmooth points, in contrast to the infinite staircases in [McDuff and Schlenk 2012; Frenkel and Müller 2012; Cristofaro-Gardiner and Kleinman 2013]. This appears to be the case for many values of b . For example, we conjecture what the function $d(a, b)$ is for all $b \geq 6$; see Conjecture 6.3.

Our proofs rely on the following remarkable theorem of Frenkel and Müller [2012]. Let $N(a, b)$ be the sequence (indexed starting at 0) of all nonnegative integer linear combinations of a and b , arranged with repetitions in nondecreasing order, and let $M(a, b)$ be the sequence whose k -th term is

$$\min\{ma + nb : (m + 1)(n + 1) \geq k + 1\},$$

where $k, m, n \in \mathbb{Z}_{\geq 0}$. Write $N(a, b) \leq M(c, d)$ if each term in the sequence $N(a, b)$ is less than or equal to the corresponding term in $M(c, d)$. Frenkel and Müller

showed that embeddings of an ellipsoid into a polydisc are completely determined by the sequences M and N :

Theorem 1.3 [Frenkel and Müller 2012]. *There is a symplectic embedding*

$$E(a, b) \xrightarrow{s} P(c, d)$$

if and only if $N(a, b) \leq M(c, d)$.

To motivate the sequences M and N , note that N is the sequence of *ECH capacities* of the symplectic ellipsoid $E(a, b)$, while M is the sequence of ECH capacities of the symplectic polydisc $P(c, d)$. The ECH capacities are a sequence of nonnegative (possibly infinite) real numbers, defined for any symplectic four-manifold, that obstruct symplectic embeddings. We will not discuss ECH capacities here; see [Hutchings 2014] for a survey. Theorem 1.3 is equivalent to the statement that the ECH capacities give sharp obstructions to embeddings of an ellipsoid into a polydisc.

2. Proof of Theorem 1.1

Weight sequences and the #-operation. We begin by describing the machinery that will be used to prove Theorem 1.1.

Let a^2 be a rational number. McDuff [2011] showed that there is a finite sequence

$$W(1, a^2) = (a_1, \dots, a_n),$$

called the *(normalized) weight sequence for a^2* , such that $E(1, a^2)$ embeds into a symplectic ellipsoid if and only if the disjoint union $\bigsqcup B(W) := \bigsqcup B(a_i)$ embeds into that ellipsoid.

To describe the weight sequence, let

$$W(a^2, 1) = (X_0^{\times \ell_0}, X_1^{\times \ell_1}, \dots, X_k^{\times \ell_k}), \tag{2-1}$$

where $X_0 > X_1 > \dots > X_k$ and $\ell_k \geq 2$. The ℓ_i are the multiplicities of the entries X_i and come from the continued fraction expansion

$$a^2 = \ell_0 + \frac{1}{\ell_1 + \frac{1}{\ell_2 + \dots + \frac{1}{\ell_k}}} := [\ell_0; \ell_1, \dots, \ell_k].$$

The entries of (2-1) are defined as

$$X_{-1} := a^2, \quad X_0 = 1, \quad X_{i+1} = X_{i-1} - \ell_i X_i \quad \text{for } i \geq 0.$$

Important properties of the weight sequence include

$$\sum_i a_i^2 = a^2, \tag{2-2}$$

$$\sum_i a_i = a^2 + 1 - \frac{1}{q}, \tag{2-3}$$

where for all i , we have $a_i \leq 1$ and $a = p/q$.

We will also make use of a helpful operation, #, as in [McDuff 2011]. Suppose s_1 and s_2 are sequences indexed with $k \in \mathbb{Z}$, starting at 0. Then,

$$(s_1 \# s_2)_k = \sup_{i+j=k} (s_1)_i + (s_2)_j.$$

A useful application of # is the following lemma:

Lemma 2.1 [McDuff 2011]. *For all $a, b > 0$, we have*

$$N(a, a) \# N(a, b) = N(a, a + b).$$

More generally, for all $\ell \geq 1$, we have

$$(\#^\ell N(a, a)) \# N(a, b) = N(a, b + \ell a).$$

This lemma together with the weight sequence and scaling implies that

$$N(1, a^2) = N(a_1, a_1) \# \cdots \# N(a_n, a_n). \tag{2-4}$$

Similar to [McDuff 2011], this machinery allows us to reduce Theorem 1.1 to a ball-packing problem.

Proof of Theorem 1.1. We begin by noting that the ECH capacities for $B(a)$ are

$$N(a, a) = (0, a, a, 2a, 2a, 2a, 3a, 3a, 3a, 3a, \dots),$$

where the terms $N_k(a, a)$ of this sequence are of the form da and for each d there are $d + 1$ entries occurring at

$$\frac{1}{2}(d^2 + d) \leq k \leq \frac{1}{2}(d^2 + 3d). \tag{2-5}$$

Similarly, for the sequence $(a/\sqrt{2b})M(1, b)$, each term $(a/\sqrt{2b})M_k(1, b)$ is of the form $d(a/\sqrt{2b})$, where

$$k \leq \frac{d^2}{4b} + \frac{(1+b)d}{2b} + \frac{b^2 - 2b + 1}{4b}. \tag{2-6}$$

By continuity, it suffices to study $d(a^2, b)$ with a^2 rational. So, we can prove that the volume obstruction is the only obstruction when $a \geq 3(b + 1)/\sqrt{2b}$ by showing that

$$N(1, a^2) \leq \frac{a}{\sqrt{2b}}M(1, b) \tag{2-7}$$

for said a -values.

By (2-5) and (2-6), it is therefore sufficient to show that

$$\sum_i d_i a_i \leq \frac{a}{\sqrt{2b}}d \tag{2-8}$$

whenever d_1, \dots, d_m, d are nonnegative integers such that

$$\sum_i (d_i^2 + d_i) \leq 2 \left(\frac{d^2}{4b} + \frac{(1+b)d}{2b} + \frac{b^2 - 2b + 1}{4b} \right). \quad (2-9)$$

We do so by considering the following cases:

Case 1: $\sum_i d_i^2 \leq d^2/(2b)$. In this case, the Cauchy–Schwarz inequality along with (2-2) implies (2-8).

Case 2: $\sum_i d_i^2 > d^2/(2b)$. This case, along with (2-9), implies

$$\sum_i d_i a_i \leq \sum_i d_i \leq \frac{(1+b)d}{b} + \frac{b^2 - 2b + 1}{2b}.$$

So, we need

$$\frac{(1+b)d}{b} + \frac{b^2 - 2b + 1}{2b} \leq \frac{a}{\sqrt{2b}} d.$$

It follows that

$$a \geq \frac{b+1}{\sqrt{2b}} \left(2 + \frac{b+1}{d} \right). \quad (2-10)$$

Now let $d = b + 1$. We see that (2-6) is equivalent to

$$k \leq b + 1 + \frac{1}{4b}.$$

It is easy to see that $N_k(1, a^2) \leq (a/\sqrt{2b})M_k(1, b)$ for all such k values. As such, we can apply $d = b + 1$ to (2-10) to get

$$a \geq \frac{3(b+1)}{\sqrt{2b}}, \quad (2-11)$$

and hence the desired result. \square

Remark 2.2. We allow $d = b + 1$ in the statement of Theorem 1.1. However, if we show $N_k(1, a^2) \leq (a/\sqrt{2b})M_k(1, b)$ for all

$$k \leq \frac{d^2}{4b} + \frac{(1+b)d}{2b} + \frac{b^2 - 2b + 1}{4b},$$

then we can use this d in (2-10) to achieve a sharper bound for a .

3. Proof of Theorem 1.2, Part I

We begin by computing $d(a, \frac{13}{2})$ on the regions where it is linear.

Nondifferentiable points and Ehrhart polynomials. We first compute d at certain values. These will eventually be the points a where $d(a, \frac{13}{2})$ is not differentiable.

Proposition 3.1. *We have*

$$\begin{aligned}
 d\left(1, \frac{13}{2}\right) &= 1, & d\left(\frac{25}{2}, \frac{13}{2}\right) &= 1, & d\left(13, \frac{13}{2}\right) &= \frac{26}{25}, \\
 d\left(\frac{351}{25}, \frac{13}{2}\right) &= \frac{26}{25}, & d\left(15, \frac{13}{2}\right) &= \frac{10}{9}, & d\left(\frac{1300}{81}, \frac{13}{2}\right) &= \frac{10}{9}, \\
 d\left(\frac{841}{52}, \frac{13}{2}\right) &= \frac{29}{26}, & d\left(17, \frac{13}{2}\right) &= \frac{34}{29}, & d\left(\frac{15028}{841}, \frac{13}{2}\right) &= \frac{34}{29}, \\
 d\left(\frac{961}{52}, \frac{13}{2}\right) &= \frac{31}{26}, & d\left(19, \frac{13}{2}\right) &= \frac{38}{31}, & d\left(\frac{18772}{961}, \frac{13}{2}\right) &= \frac{38}{31}, \\
 d\left(\frac{1089}{52}, \frac{13}{2}\right) &= \frac{33}{26}, & d\left(21, \frac{13}{2}\right) &= \frac{42}{33}, & d\left(\frac{2548}{121}, \frac{13}{2}\right) &= \frac{42}{33}.
 \end{aligned}$$

To prove the proposition, the main difficulty comes from the fact that applying Theorem 1.3 in principle requires checking infinitely many ECH capacities. Our strategy for overcoming this difficulty is to study the growth rate of the terms in the sequences M and N . We will find that in every case needed to prove Proposition 3.1, one can bound these growth rates to conclude that only finitely many terms in the sequences need to be checked. This is then easily done by computer. The details are as follows:

Proof. Step 1: For the sequence $N(a, b)$, let $k(a, b, t)$ be the largest k such that $N_k(a, b) \leq t$. Similarly, for the sequence $M(c, d)$, let $l(c, d, t)$ be the largest l such that $M_l(c, d) \leq t$. To show that $E(a, b) \xrightarrow{s} P(c, d)$, by Theorem 1.3, we just have to show that for all t , we have $k(a, b, t) \geq l(c, d, t)$.

Step 2: We can estimate $k(a, b, t)$ by applying the following proposition:

Proposition 3.2. *If a, b, r , and t are all positive integers, then*

$$\begin{aligned}
 k\left(\frac{a}{r}, \frac{b}{r}, t\right) &= \frac{1}{2ab}(tr)^2 + \frac{1}{2}(tr)\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{ab}\right) + \frac{1}{4}\left(1 + \frac{1}{a} + \frac{1}{b}\right) + \frac{1}{12}\left(\frac{a}{b} + \frac{b}{a} + \frac{1}{ab}\right) \\
 &\quad + \frac{1}{a} \sum_{j=1}^{a-1} \frac{\xi_a^{j(-tr)}}{(1-\xi_a^{jb})(1-\xi_a^j)} + \frac{1}{b} \sum_{l=1}^{b-1} \frac{\xi_b^{l(-tr)}}{(1-\xi_b^{la})(1-\xi_b^l)}, \tag{3-1}
 \end{aligned}$$

where $\xi_d = e^{2\pi i/d}$.

Proof. The number of terms in $N(a/r, b/r)$ that are less than t is the same as the number of lattice points (m, n) in the triangle bounded by the positive x - and y -axes and the line $x(a/r) + y(b/r) \leq t$. For integral t , this number can be computed by applying the theory of ‘‘Ehrhart polynomials’’. Proposition 3.2 follows by applying [Beck and Robins 2007, Theorem 2.10]. \square

We will be most interested in this proposition in the case where $a = r$. Note that by the last two terms of the formula in Proposition 3.2, we have $k(a/r, b/r, t)$ is a periodic polynomial with period ab .

We also need an argument to account for the fact that Proposition 3.2 is only for integral t , whereas the argument in Step 1 involves real t . To account for this,

we use an asymptotic argument. Specifically, for $E(1, a/r)$, with $a, r \in \mathbb{Z}_{\geq 1}$, we bound the right-hand side of (3-1) from below by taking the floor function of t . It is convenient for our argument to further bound this expression from below by

$$\frac{c_1}{r^2}(rt - 1)^2 + \frac{c_2}{r}(rt - 1) + c_3, \tag{3-2}$$

where the c_i are the coefficients of the right-hand side of (3-1) that do not involve t or r .

This is the lower bound that we will use for $k(1, a/r, t)$.

Step 3: To get an upper bound $l(c, d, t)$ for $M(c, d)$, recall that

$$M_l(c, d) = \min\{cm + dn : (m + 1)(n + 1) \geq l + 1\}.$$

For $cm + dn = t$, we solve for m in terms of n and find

$$\left(\frac{t - dn}{c} + 1\right)(n + 1) - 1 \geq l.$$

Considering $m, n \in \mathbb{R}$, we can take the derivative of the left side of the inequality with respect to n and then set the expression equal to 0 to maximize it. We do the same with m to obtain

$$\left(\frac{t}{2d} + \frac{c}{2d} + \frac{1}{2}\right)\left(\frac{t}{2c} + \frac{d}{2c} + \frac{1}{2}\right) - 1 \geq l.$$

By simplifying, we get that an upper bound for l is

$$l(c, d, t) = \frac{t^2}{4cd} + \frac{(c + d)t}{2cd} + \frac{(c - d)^2}{4cd}. \tag{3-3}$$

Our strategy now is to check that for each point in Proposition 3.1, we have $k(a, b, t) \geq l(c, d, t)$ asymptotically in t for the corresponding (a, b, c, d) . From there, we can check that for a sufficient number of terms, $N(1, a) \leq M(\lambda, \lambda b)$.

Step 4: Since the rest of the proof amounts to computation, it is best summarized by Table 1. In the table, k_{t^2} and l_{t^2} denote the coefficients of the quadratic terms in the upper and lower bounds from Steps 2 and 3, while k_t and l_t denote the corresponding coefficients of the linear terms.

The t -column gives a sufficient number to check up to before the asymptotic bounds from the previous three steps are enough. Note that if k_{t^2} and l_{t^2} in any row are equal, then linear coefficients, k_t and l_t , are used to make an asymptotic argument; this explains the appearance of the ‘‘N/A’’s in the table. It is simple to check by computer that the relevant N and M sequences in each row satisfy $N \leq M$ once one knows that the problem only has to be checked up to the t in the t -column.

The rightmost column of Table 1 gives an ECH capacity that shows that one cannot shrink λ further, i.e., the claimed embeddings are actually sharp. \square

$E(1, a) \xrightarrow{s} P(\lambda, \lambda b)$	k_{t_2}	l_{t_2}	k_t	l_t	t	ECH obstruction
$E(1, \frac{25}{2}) \xrightarrow{s} P(1, \frac{13}{2})$	$\frac{1}{25}$	$\frac{1}{26}$	N/A	N/A	51	1
$E(1, 13) \xrightarrow{s} P(\frac{26}{25}, \frac{169}{25})$	$\frac{1}{26}$	$\frac{625}{17576}$	N/A	N/A	33	13
$E(1, \frac{351}{25}) \xrightarrow{s} P(\frac{26}{25}, \frac{169}{25})$	$\frac{25}{702}$	$\frac{625}{17576}$	N/A	N/A	522	13
$E(1, 15) \xrightarrow{s} P(\frac{10}{9}, \frac{65}{9})$	$\frac{1}{30}$	$\frac{81}{2600}$	N/A	N/A	29	15
$E(1, \frac{1300}{81}) \xrightarrow{s} P(\frac{10}{9}, \frac{65}{9})$	$\frac{81}{2600}$	$\frac{81}{2600}$	$\frac{691}{1300}$	$\frac{27}{52}$	272	15
$E(1, \frac{841}{52}) \xrightarrow{s} P(\frac{29}{26}, \frac{29}{4})$	$\frac{26}{841}$	$\frac{26}{841}$	$\frac{447}{841}$	$\frac{15}{29}$	122	17
$E(1, 17) \xrightarrow{s} P(\frac{34}{29}, \frac{221}{29})$	$\frac{1}{34}$	$\frac{841}{30056}$	N/A	N/A	27	17
$E(1, \frac{15028}{841}) \xrightarrow{s} P(\frac{34}{29}, \frac{221}{29})$	$\frac{841}{30056}$	$\frac{841}{30056}$	$\frac{7935}{15028}$	$\frac{435}{884}$	32	17
$E(1, \frac{961}{52}) \xrightarrow{s} P(\frac{31}{26}, \frac{31}{4})$	$\frac{26}{961}$	$\frac{26}{961}$	$\frac{507}{961}$	$\frac{15}{31}$	23	19
$E(1, 19) \xrightarrow{s} P(\frac{38}{31}, \frac{247}{31})$	$\frac{1}{38}$	$\frac{961}{37544}$	N/A	N/A	7	19
$E(1, \frac{18772}{961}) \xrightarrow{s} P(\frac{38}{31}, \frac{247}{31})$	$\frac{961}{37544}$	$\frac{961}{37544}$	$\frac{759}{1444}$	$\frac{465}{988}$	28	19
$E(1, \frac{1089}{52}) \xrightarrow{s} P(\frac{33}{26}, \frac{33}{4})$	$\frac{26}{1089}$	$\frac{26}{1089}$	$\frac{571}{1089}$	$\frac{15}{33}$	14	21
$E(1, 21) \xrightarrow{s} P(\frac{42}{33}, \frac{273}{33})$	$\frac{1}{42}$	$\frac{121}{5096}$	N/A	N/A	26	21
$E(1, \frac{2548}{121}) \xrightarrow{s} P(\frac{42}{33}, \frac{273}{33})$	$\frac{121}{5096}$	$\frac{121}{5096}$	$\frac{1335}{2548}$	$\frac{165}{364}$	41	21

Table 1. The computations from Step 4 of the proof of Proposition 3.1.

The linear steps. Given the computations from the previous section, the computation of $d(a, \frac{13}{2})$ for all the “linear steps”, i.e., those portions of the graph of d for which d is linear, is straightforward. Indeed, we have the following two lemmas:

Lemma 3.3. *For fixed b , the function $d(a, b)$ is monotonically nondecreasing.*

Proof. This follows from the fact that $E(1, a) \xrightarrow{s} E(1, a')$ if $a \leq a'$. □

Lemma 3.4 (subscaling). $d(\lambda a, b) \leq \lambda d(a, b)$.

Proof. This follows from the fact that $E(1, \lambda a) \xrightarrow{s} E(\lambda, \lambda a)$ for $\lambda \geq 1$. □

By monotonicity, we know that $d(a, \frac{13}{2})$ is constant on the intervals

$$[1, \frac{25}{2}], \quad [13, \frac{351}{25}], \quad [15, \frac{1300}{81}], \quad [17, \frac{15028}{841}], \quad [19, \frac{18772}{961}], \quad [21, \frac{2548}{121}].$$

We now explain why for $k \in \mathbb{Z}$, with $0 \leq k \leq 4$, we have

$$d(a, \frac{13}{2}) = \frac{2a}{25 + 2k} \quad \text{for } a \in [\alpha_k, 13 + 2k],$$

where $\alpha_0 = \frac{25}{2}, \alpha_1 = \frac{351}{25}, \alpha_2 = \frac{841}{52}, \alpha_3 = \frac{961}{52}$, and $\alpha_4 = \frac{1089}{52}$.

Given the critical points we have determined, along with the subscaling lemma, we have $2a/(25 + 2k)$ as an upper bound for $d(a, \frac{13}{2})$ on the above intervals.

Intervals on which $d(a, \frac{13}{2})$ is linear. We also know that

$$d(a, \frac{13}{2}) = \sup \left\{ \frac{N_x(1, a)}{M_x(1, \frac{13}{2})} : x \in \mathbb{N} \right\} \geq \frac{N_l(1, a)}{M_l(1, \frac{13}{2})} \quad \text{for any } l.$$

Here is a representative example of our method:

Example 3.5. To illustrate how this can give us a suitable lower bound, consider the case where $x = 13$:

$$\sup \left\{ \frac{N_x(1, a)}{M_x(1, \frac{13}{2})} : x \in \mathbb{N} \right\} \geq \frac{N_{13}(1, a)}{M_{13}(1, \frac{13}{2})} = \frac{2a}{25} \quad \text{for } a \in [\frac{25}{2}, 13].$$

This lower bound equals the upper bound given by Lemma 3.4, so we have proven Theorem 1.2 for $a \in [\frac{25}{2}, 13]$.

The general method is similar: given $a \in [\alpha_k, 13 + 2k]$, we can find an l such that

$$\frac{N_l(1, a)}{M_l(1, \frac{13}{2})} = \frac{2a}{25 + 2k}.$$

Such obstructing values of l are given in the following table:

k	$\frac{2}{25+2k}$	l
0	$\frac{2}{25}$	13
1	$\frac{2}{27}$	15
2	$\frac{2}{29}$	17
3	$\frac{2}{31}$	19
4	$\frac{2}{33}$	21

Given $a \in [\alpha_k, 13 + 2k]$ for each integer $k \in [0, 4]$, we have found that the upper and lower bounds of $d(a, \frac{13}{2})$ equal $2a/(25 + 2k)$. Thus, we have proven our claim for these intervals.

4. Proof of Theorem 1.2, Part II

To complete the proof of Theorem 1.2, we need to show that aside from the linear steps described in the previous section, the graph of $d(a, \frac{13}{2})$ is equal to the graph of the volume obstruction. To do this, we adapt some of the ideas from [McDuff and Schlenk 2012] in a purely combinatorial way. This will be needed to complete

the proof of Theorem 1.2. Our combinatorial perspective on the techniques from [McDuff and Schlenk 2012] borrows many ideas from [McDuff 2011].

Preliminaries. This section collects the main combinatorial machinery that will be used to complete the proof. The basic idea behind our proof will be to reduce to a ball-packing problem, as in the proof of Theorem 1.1. The machinery we develop here will be useful for approaching this ball-packing problem.

We begin with two definitions:

Definition 4.1. Let $\text{Cr}(d, d_i) = (d', d'_i)$, where $d' = 2d - d_1 - d_2 - d_3$, $d'_i = d - d_j - d_k$ for $i, j, k = 1, 2, 3$ and $d'_i = d_i$ for all $i \geq 4$. We call Cr the *Cremona transform*.

Definition 4.2. We say $(d, d_i) \in \mathbb{R}^{1+n}$ is

- (i) *positive* if $d, d_i \geq 0$ for all i ,
- (ii) *ordered* if $d_i, d_{i+1} \neq 0$ implies $d_i \geq d_{i+1}$ and $d_i \neq 0, d_j = 0$ implies $i < j$,
- (iii) *reduced* if positive, ordered, and $d \geq d_1 + d_2 + d_3$.

Remark 4.3. It will be important to note that $\text{Cr}(\text{Cr}(d, d_i)) = (d, d_i)$.

We now define a product analogous to the intersection product in [McDuff and Schlenk 2012]:

Definition 4.4.
$$(x, x_i) \cdot (y, y_i) = xy - \sum_i x_i y_i.$$

We also define a vector $-K \in \mathbb{R}^{1+n}$ that is motivated by the standard anticanonical divisor in the M -fold blow up of $\mathbb{C}P^2$.

Definition 4.5.
$$-K = (3, 1, 1, \dots, 1).$$

The following is a combinatorial analogue of “positivity of intersections” that will be useful:

Lemma 4.6. *If (x, x_i) is reduced, (d, d_i) is positive, $-K \cdot (d, d_i) \geq 0$, and $d \geq \max(d_i)$, then $(x, x_i) \cdot (d, d_i) \geq 0$.*

Proof. Let (d', d'_i) be the vector obtained from ordering d_i . As

$$(x, x_i) \cdot (d, d_i) \geq (x, x_i) \cdot (d', d'_i),$$

we can assume without loss of generality that (d, d_i) is ordered. If $x_3 = 0$ then $x_i = 0$ for $i \geq 3$ and

$$(x, x_i) \cdot (d, d_i) = xd - x_1 d_1 - x_2 d_2.$$

As $d \geq \max(d_i)$, we know that this expression is greater than or equal to

$$(x - x_1 - x_2)d.$$

As (x, x_i) is reduced, this is greater than or equal to 0.

We now assume without loss of generality that $x_3 = 1$. Hence, $x_i \leq 1$ for $i \geq 3$. Let $e_1 = x_1 - 1$ and $e_2 = x_2 - 1$. Then

$$xd \geq (3 + e_1 + e_2)d$$

as (x, x_i) is reduced. This expression is equal to

$$3d + de_1 + de_2.$$

As $d \geq d_1, d_2$, we now have the following chain of inequalities:

$$\begin{aligned} 3d + de_1 + de_2 &\geq 3d + d_1e_1 + d_2e_2 \geq \sum_i d_i + d_1e_1 + d_2e_2 \\ &= d_1x + d_2x + \sum_{i \geq 3} d_i \geq d_1x_1 + d_2x_2 + \sum_{i \geq 3} x_i d_i = \sum_i d_i x_i. \quad \square \end{aligned}$$

In [McDuff and Schlenk 2012], Cremona transformations preserve the intersection product. Here we prove an analogous result.

Lemma 4.7. $\text{Cr}(x, x_i) \cdot \text{Cr}(y, y_i) = (x, x_i) \cdot (y, y_i)$.

Proof.

$$\begin{aligned} \text{Cr}(x, x_i) \cdot \text{Cr}(y, y_i) &= x' y' - \sum_i x'_i y'_i \\ &= (2x - x_1 - x_2 - x_3)(2y - y_1 - y_2 - y_3) - (x - x_2 - x_3)(y - y_2 - y_3) \\ &\quad - (x - x_1 - x_3)(y - y_1 - y_3) - (x - x_2 - x_3)(y - y_2 - y_3) - \sum_{i > 3} x_i y_i \\ &= xy - x_1 y_1 - x_2 y_2 - x_3 y_3 - \sum_{i > 3} x_i y_i \\ &= xy - \sum_i x_i y_i \\ &= (x, x_i) \cdot (y, y_i). \quad \square \end{aligned}$$

The following sets will also be useful:

Definition 4.8. $F = \{(d, d_i) : (d, d_i) \cdot (-K + (d, d_i)) \geq 0, d, d_i \in \mathbb{Z}\}$.

Definition 4.9. $F^+ = \{(d, d_i) : (d, d_i) \in F, d, d_i \geq 0\}$.

Definition 4.10. $E = \{(d, d_i) : (d, d_i) \cdot (d, d_i) \geq -1, -K \cdot (d, d_i) = 1, d, d_i \in \mathbb{Z}\}$.

Remark 4.11. Observe $\text{Cr}(F) \subset F$ and $\text{Cr}(E) \subset E$. Additionally, F, F^+ , and E are invariant under permutations of d_i .

Remark 4.12. Note that $(0, -1, 0, \dots, 0) \in E$.

Definition 4.13. Let C be the set of (x, x_i) such that $x, x_i \in \mathbb{Z}$ and

- (a) $(x, x_i) \cdot (x, x_i) \geq 0$,
- (b) $(x, x_i) \cdot (d, d_i) \geq 0$ for all $(d, d_i) \in E$.

Both Li and Li [2002] and McDuff and Schlenk [2012] have found that compositions of Cremona transformations and permutations can reduce certain classes. Here we prove a combinatorial version of those lemmas.

Lemma 4.14. *If $(x, x_i) \in C$ then by a sequence of Cremona transforms and permutations of x_i , we can transform (x, x_i) to (x', x'_i) , where (x', x'_i) is reduced.*

Proof. We begin with some helpful results:

Sublemma 4.15. $\text{Cr}(C) \subset C$.

Proof. The fact that Cr preserves (a) follows from the fact that Cr preserves the intersection product. To complete the sublemma, note that if $(d, d_i) \in E$, then

$$\text{Cr}(x, x_i) \cdot (d, d_i) = \text{Cr}^2(x, x_i) \cdot (d', d'_i) = (x, x_i) \cdot (d', d'_i) \geq 0 \quad \text{as } (d', d'_i) \in E. \quad \square$$

Sublemma 4.16. *If P is some permutation, $P(C) \subset C$.*

Proof. If $(d, d_i) \in E$, then

$$P(x, x_i) \cdot (d, d_i) = (x, x_i) \cdot P^{-1}(d, d_i) \quad \text{as } P^{-1}(E) \subset E. \quad \square$$

Sublemma 4.17. *If $(x, x_i) \in C$, then $x, x_i \geq 0$.*

Proof. If $d_i = (-\delta_{ij})$ and $(0, d_i) \in E$ then we have $j \leq \text{length}(d_i)$ for all j . So, $(x, x_i) \cdot (0, d_i) = x_j \geq 0$. We also have $(x, x_i) \cdot (1, 1, 1, 0, 0, \dots, 0) = x - x_1 - x_2 \geq 0$. As $x_1, x_2 \geq 0$, this implies that $x \geq 0$. \square

Let oCr denote the transformation Cr followed by ordering the d_i . Fix $(x, x_i) \in C$. Let $(x^k, x_i^k) = \text{oCr}^k(x, x_i)$. Let $\alpha(k) = x^k - x_1^k - x_2^k - x_3^k$. It suffices to show $\alpha(k) \geq 0$ for some k . Assume not. Then $\alpha(k) \leq -1$ for all k . By Sublemmas 4.15 and 4.16, $\text{oCr}(C) \subset C$. For $k \geq 1$,

$$x^k = x^{k-1} + \alpha(k-1) \leq x^{k-1} - 1.$$

Thus, there exists k such that $x^k < 0$. This contradicts Sublemma 4.17, completing the proof that we may reduce (x, x_i) . \square

We now prove Lemma 4.18, a result analogous to [McDuff and Schlenk 2012, Proposition 1.2.12(i)].

Lemma 4.18. *If $(x, x_i) \in C$ then $(x, x_i) \cdot (d, d_i) \geq 0$ for all $(d, d_i) \in F$.*

Proof. By Lemma 4.14 there exists A , a composition of Cr and permutations, such that $A(x, x_i) = (x', x'_i)$ with (x', x'_i) reduced. For $(d, d_i) \in F$, let $A(d, d_i) = (d', d'_i) \in F$. So,

$$(x, x_i) \cdot (d, d_i) = A(x, x_i) \cdot A(d, d_i) = (x', x'_i) \cdot (d', d'_i).$$

Let $e = d$, $e_i = d_i$ if $d_i > 0$ and $e_i = 0$ if $d_i \leq 0$. We note $(e, e_i) \in F$ and

$$(x', x'_i) \cdot (d', d'_i) \geq (x', x'_i) \cdot (e, e_i).$$

If $(e, e_i) \cdot (e, e_i) \geq 0$ then the Cauchy–Schwarz inequality shows $(x', x'_i) \cdot (e, e_i) \geq 0$. Otherwise, $(e, e_i) \cdot (-K) \geq 0$. Then

$$\sum_i e_i^2 + e_i \leq e^2 + 3e$$

implies $e \geq e_i$, so Lemma 4.6 shows $(x', x'_i) \cdot (e, e_i) \geq 0$. □

Remark 4.19. By scaling, Lemma 4.18 extends to (x, x_i) that satisfy (a) and (b) of Definition 4.13 with $x, x_i \in \mathbb{Q}$.

A key lemma. We now use the combinatorial machinery from the previous section, together with a reduction to the ball-packing problem, to prove the key lemma needed to complete the proof of Theorem 1.2; see part (iii) of Lemma 4.24 below.

To reduce to a ball-packing problem, note that [Frenkel and Müller 2012, Proposition 1.4] states that for rational a , we have that

$$E(1, a) \xrightarrow{S} P(\lambda, c\lambda)$$

if and only if

$$E(1, a) \sqcup B(\lambda) \sqcup B(c\lambda) \xrightarrow{S} B((1+c)\lambda), \tag{4-1}$$

where \sqcup denotes disjoint union. Since, as explained in [Hutchings 2014], one can compute the ECH capacities of the disjoint union in terms of the $\#$ -operation, we know that the embedding in (4-1) exists if and only if

$$N(1, a) \# N(\lambda, \lambda) \# N(c\lambda, c\lambda) \leq N((1+c)\lambda, (1+c)\lambda). \tag{4-2}$$

For the rest of the proof of Theorem 1.2, we are looking at intervals for a on which the graph of d is equal to the volume obstruction; we therefore want to show that (4-2) holds with $\lambda = \sqrt{a}/(2c)$ (of course, for our proof one can specify $c = \frac{13}{2}$, but we state things here in slightly greater generality). By an argument analogous to the argument used in the proof of Theorem 1.1, it is sufficient to show

$$\left(\sum_i d_i^2 + d_i \right) + e_1^2 + e_1 + e_2^2 + e_2 \leq d^2 + 3d$$

implies

$$\left(\sum_i a_i d_i \right) + c\lambda e_1 + \lambda e_2 \leq (1+c)\lambda d$$

for all nonnegative integers d, d_i, e_1, e_2 . Let $m_1 = e_1, m_2 = e_2$ and $m_i = d_{i-2}$ for $i \geq 3$ and let $w_1(a) = c\lambda, w_2(a) = \lambda$ and $w_i(a) = a_{i-2}$ for $i \geq 3$. Hence, it is enough to show

$$\sum_i m_i^2 + m_i \leq d^2 + 3d$$

implies

$$m \cdot w(a) \leq (1+c)\lambda d. \tag{4-3}$$

Let $\mu(d; m)(a) = (m \cdot w(a))/d$. Then (4-3) is equivalent to $\mu(d; m)(a) \leq (1+c)\lambda$. By Lemma 4.18, it is sufficient to check the case

$$\sum_i m_i^2 = d^2 + 1, \quad (4-4)$$

$$\sum_i m_i = 3d - 1. \quad (4-5)$$

Let E be the set of $(d; m)$ satisfying (4-4) and (4-5) with d, m_i nonnegative integers. Define ε by

$$m = \frac{d}{(1+c)\lambda} w(a) + \varepsilon.$$

We now have a series of lemmas, culminating in the key lemma, Lemma 4.24.

Lemma 4.20. *For $(d; m) \in E$, we have:*

- (i) $\mu(d; m)(a) \leq (1+c)\lambda \sqrt{1 + \frac{1}{d^2}}$.
- (ii) $\mu(d; m)(a) > (1+c)\lambda$ if and only if $\varepsilon \cdot w > 0$.
- (iii) $\mu(d; m)(a) > (1+c)\lambda$ implies $\sum_i \varepsilon_i^2 < 1$.
- (iv) Let $y(a) = a + 1 - 2(1+c)\lambda$. Then

$$-\sum_i \varepsilon_i = 1 + \frac{d}{(1+c)\lambda} \left(y(a) - \frac{1}{q} \right),$$

where $a = p/q$.

Proof. Part (i) follows from $\sum_i w_i^2 = c^2\lambda^2 + \lambda^2 + \sum_i a_i^2 = (1+c)^2\lambda^2$ and the Cauchy–Schwarz inequality. To prove (ii), note

$$\begin{aligned} \varepsilon \cdot w &= m \cdot w - \frac{d}{(1+c)\lambda} w \cdot w \\ &= d \left(\frac{m \cdot w}{d} - (1+c)\lambda \right) \\ &= d(\mu(d; m)(a) - (1+c)\lambda). \end{aligned}$$

To prove (iii), note

$$\begin{aligned} \sum_i \varepsilon_i^2 &= \varepsilon \cdot \varepsilon = m \cdot m + \frac{d^2}{(1+c)^2\lambda^2} w \cdot w - \frac{2d}{(1+c)\lambda} m \cdot w \\ &= 1 + d^2 \left(2 - \frac{2}{(1+c)\lambda} \frac{m \cdot w}{d} \right) \\ &< 1 \quad \text{if } \mu(d; m)(a) > (1+c)\lambda. \end{aligned}$$

To prove (iv), note

$$\begin{aligned}
 -\sum_i \varepsilon_i &= \frac{d}{(1+c)\lambda} \sum_i w_i - \sum_i m_i \\
 &= \frac{d}{(1+c)\lambda} \left(a + 1 - \frac{1}{q} + c\lambda + \lambda \right) - 3d - 1 \\
 &= 1 + \frac{d}{(1+c)\lambda} \left(a + 1 - \frac{1}{q} - 2(1+c)\lambda \right). \quad \square
 \end{aligned}$$

Lemma 4.21. *Let $(d; m) \in E$ and suppose that I is the maximal nonempty open interval such that $\mu(d; m)(a) > (1+c)\lambda$ for all $a \in I$. Then there exists a unique $a_0 \in I$ such that $l(a_0) = l(m)$, where $l(a_0)$ is the length of $w_i(a)$ and $l(m)$ is the number of nonzero terms in m . Furthermore, $l(a) \geq l(m)$ for all $a \in I$.*

Proof. We adapt the proof of Lemma 2.1.3 in [McDuff and Schlenk 2012]. For $i \geq 3$, $w_i(a)$ is piecewise linear and is linear on open intervals that do not contain an element a' with length $l(a') \leq i$. Therefore, if $l(a) > l(m)$ for all $a \in I$,

$$\mu(d; m)(a) - \frac{c\lambda m_1 + \lambda m_2}{d}$$

is linear on I . This is impossible as $c\lambda(1 - m_1/d) + \lambda(1 - m_2/d)$ is concave and I is bounded. Thus there exists $a_0 \in I$ with $l(a_0) \leq l(m)$. If $l(a) < l(m)$ then $\sum_{i \leq l(a)} m_i^2 < d^2 + 1$, which implies

$$m \cdot w \leq \|w\| \sqrt{\sum_{i \leq l(a)} m_i^2} \leq d \|w\| = (1+c)\lambda d,$$

which is impossible for $a \in I$. The proof of uniqueness is the same as in [McDuff and Schlenk 2012, Lemma. 2.1.3]. □

Lemma 4.22. *Let $(d; m)$ be in E with $\mu(d; m)(a) > (1+c)\lambda$ for some a . Let $J = k, \dots, k+s-1$ be a block of $s \geq 2$ consecutive integers such that $w_i(a)$ is constant for $i \in J$. Then:*

(i) *One of the following holds:*

- $m_k = \dots = m_{k+s-1}$.
- $m_k = \dots = m_{k+s-2} = m_{k+s-1} + 1$.
- $m_k - 1 = m_{k+1} = \dots = m_{k+s-1}$.

(ii) *There is at most one block of length $s \geq 2$ on which the m_i are not all equal.*

(iii) *If there is a block J of length $s \geq 2$ on which the m_i are not all equal then*

$$\sum_{i \in J} \varepsilon_i^2 \geq \frac{s-1}{s}.$$

Proof. See the proof of [McDuff and Schlenk 2012, Lemma. 2.1.7]. McDuff and Schlenk consider the case of embedding an ellipsoid into a ball, but their proof generalizes without change to our situation. \square

Lemma 4.23. *Let $(d; m) \in E$ be such that $\mu(d; m) > (1 + c)\lambda$ for some a with $l(a) = l(m) = M$. Let w_{k+1}, \dots, w_{k+s} be a block, but not the first block, of $w(a)$ (the first two terms of $w(a)$ are not considered to be part of any block).*

(i) *If this block is not the last block, then*

$$|m_k - (m_{k+1} + \dots + m_{k+s} + m_{k+s+1})| < \sqrt{s + 2}.$$

If this block is the last block, then

$$|m_k - (m_{k+1} + \dots + m_{k+s})| < \sqrt{s + 1}.$$

(ii) *It is always true that*

$$m_k - \sum_{i=k+1}^M m_i < \sqrt{M - k + 1}.$$

Proof. This is similar to the proof of Lemma 4.22; see the proof of [McDuff and Schlenk 2012, Lemma 2.1.8], which generalizes without change to our situation. \square

Lemma 4.24. *Assume that $(d; m) \in E$ and $\mu(d; m)(a) > (1 + c)\lambda$ for some a with $l(a) = l(m)$. Assume further that $y(a) > 1/q$. Let*

$$v_M = \frac{d}{q(1 + c)}\lambda$$

and let $L = l(m)$. Then:

(i) $|\sum_i \varepsilon_i| \leq \sqrt{L}.$

(ii) $v_M > \frac{1}{3}.$

(iii) *Let $\delta = y(a) - 1/q > 0$. Then*

$$d \leq \frac{(1 + c)\lambda}{\delta}(\sqrt{L} - 1) \leq \frac{(1 + c)\lambda}{\delta}(\sqrt{q + [a] + 2} - 1)$$

and $\sqrt{q + [a] + 2} \geq 1 + \delta v_M q.$

Proof. Part (i) follows from $\sum_i \varepsilon_i^2 < 1$. Part (ii) follows from the same argument as [McDuff and Schlenk 2012, Lemma 5.1.2]. From [McDuff and Schlenk 2012, Sublemma 5.1.1], we have $q + [a] + 2 \geq L$, so Lemma 4.20 implies

$$\sqrt{q + [a] + 2} \geq \sqrt{L} \geq 1 + \frac{d}{(1 + c)\lambda} \left(y(a) - \frac{1}{q} \right) = 1 + \frac{d}{(1 + c)\lambda} \delta = 1 + q v_M \delta.$$

This also shows

$$d \leq \frac{(1 + c)\lambda}{\delta}(\sqrt{q + [a] + 2} - 1). \quad \square$$

5. Proof of Theorem 1.2, Part III

With the Lemma 4.24 now shown, we can complete the proof of Theorem 1.2. We explain the computation on various intervals separately.

$[\frac{1300}{81}, \frac{841}{52}]$. We now wish to prove that $d(a, \frac{13}{2}) = \sqrt{a/13}$ for $a \in [\frac{1300}{81}, \frac{841}{52}]$. Previously, we proved

$$\left(\frac{1300}{81}, \frac{13}{2}\right) = \frac{10}{9} \quad \text{and} \quad d\left(\frac{841}{52}, \frac{13}{2}\right) = \frac{29}{26}.$$

If $d(a, \frac{13}{2})$ is not equal to $\sqrt{a/13}$ on the interval $[\frac{1300}{81}, \frac{841}{52}]$, there exists $(d; m) \in E$ such that

$$\mu(d; m)(a) > 7.5\lambda \quad \text{for some } a \in \left[\frac{1300}{81}, \frac{841}{52}\right].$$

So, Lemma 4.24 shows that there exists a_0 in $[\frac{1300}{81}, \frac{841}{52}]$ with $\mu(d; m)(a_0) > 7.5\lambda$ and $l(a_0) = l(m)$. Let $a_0 = p/q = 16 + p'/q$. As $16 < a_0 < 16 + \frac{1}{5}$, we know $q \geq 5$. For $a_0 \in [\frac{1300}{81}, \frac{841}{52}]$ and $q \geq 5$, we know

$$\delta \geq \frac{1300}{81} + 1 - 15\sqrt{\frac{1300}{81 \cdot 13}} - \frac{1}{q} \geq \frac{31}{81} - \frac{1}{q}.$$

Thus, Lemma 4.24 shows

$$\sqrt{q+18} \geq 1 + \left(\frac{31}{81} - \frac{1}{q}\right)\frac{q}{3}.$$

Hence, $q \leq 67$.

We also note that for $\frac{1300}{81} < a_0 < \frac{841}{52}$ and $q \geq 5$, we have

$$\lambda \leq \sqrt{\frac{841}{52 \cdot 13}} = \frac{29}{26} \quad \text{and} \quad \delta \geq \frac{31}{81} - \frac{1}{q} \geq \frac{74}{405}.$$

Thus, Lemma 4.24 shows

$$d \leq \frac{7.5 \cdot \frac{29}{26}}{\frac{74}{405}} (\sqrt{85} - 1) < 377.$$

Using Mathematica we can reduce the possibilities for $(d; m)$ to 38 candidates. We can then use Lemma 4.23 to reduce these 38 cases to 11 possible candidates, which can easily verified to not be obstructive by simple calculations.

$[\frac{15028}{841}, \frac{961}{52}]$. We now will show $d(a, \frac{13}{2}) = \sqrt{a/13}$ for $a \in [\frac{15028}{841}, \frac{961}{52}]$. Previously, we proved

$$d\left(\frac{15028}{841}, \frac{13}{2}\right) = \frac{34}{29} \quad \text{and} \quad d\left(\frac{961}{52}, \frac{13}{2}\right) = \frac{31}{26}.$$

If $d(a, \frac{13}{2})$ is not equal to $\sqrt{a/13}$ on the interval $[\frac{15028}{841}, \frac{961}{52}]$, then there exists $(d; m) \in E$ such that

$$\mu(d; m)(a) > 7.5\lambda \quad \text{for some } a \in \left[\frac{15028}{841}, \frac{961}{52}\right].$$

Then Lemma 4.24 shows that there exists $a_0 \in [\frac{15028}{841}, \frac{961}{52}]$ with $\mu(d, m)(a_0) > 7.5\lambda$ and $l(a_0) = l(m)$. Let $a_0 = p/q$ with $\gcd(p, q) = 1$. For $a_0 \in [\frac{15028}{841}, \frac{961}{52}]$, we know

$$\delta \geq \frac{15028}{841} + 1 - 15\sqrt{\frac{15028}{841 \cdot 13}} - \frac{1}{q} = \frac{1079}{841} - \frac{1}{q}.$$

Thus, Lemma 4.24 shows

$$\sqrt{q+19} \geq 1 + \left(\frac{1079}{841} - \frac{1}{q}\right)\frac{q}{3}.$$

Hence, $q \leq 11$. We can then verify these cases directly using Mathematica, which by simple calculations can be verified not to be obstructive.

$[\frac{18772}{961}, \frac{1089}{52}]$. We will now show $d(a, \frac{13}{2}) = \sqrt{a/13}$ for $a \in [\frac{18772}{961}, \frac{1089}{52}]$. Previously, we proved

$$d\left(\frac{18772}{961}, \frac{13}{2}\right) = \frac{38}{31} \quad \text{and} \quad d\left(\frac{1089}{52}, \frac{13}{2}\right) = \frac{33}{26}.$$

If $d(a, \frac{13}{2})$ is not equal to $\sqrt{a/13}$ on the interval $[\frac{18772}{961}, \frac{1089}{52}]$, then there exists $(d; m) \in E$ such that

$$\mu(d; m)(a) > 7.5\lambda \quad \text{for some } a \in \left[\frac{18772}{961}, \frac{1089}{52}\right].$$

Then Lemma 4.24 shows that there exists $a_0 \in [\frac{18772}{961}, \frac{1089}{52}]$ with $\mu(d, m)(a_0) > 7.5\lambda$ and $l(a_0) = l(m)$. Let $a_0 = p/q$ with $\gcd(p, q) = 1$. For $a_0 \in [\frac{18772}{961}, \frac{1089}{52}]$, we know

$$\delta \geq \frac{18772}{961} + 1 - 15\sqrt{\frac{18772}{961 \cdot 13}} - \frac{1}{q} = \frac{2063}{961} - \frac{1}{q}.$$

Thus, Lemma 4.24 shows

$$\sqrt{q+21} \geq 1 + \left(\frac{2063}{961} - \frac{1}{q}\right)\frac{q}{3}.$$

Hence, $q \leq 6$. We can then verify these cases directly using Mathematica to check these cases and we find no obstructions.

$[\frac{2548}{121}, 27]$. For $a \in [\frac{2548}{121}, 27]$, we have

$$\sqrt{q+29} \geq \sqrt{q+[a]+2} \quad \text{and} \quad \delta \geq 21 - 15\sqrt{\frac{21}{13}}.$$

Hence, Lemma 4.24 implies

$$\sqrt{q+29} \geq 1 + \left(21 - 15\sqrt{\frac{21}{13}}\right)\frac{q}{3},$$

which implies $q < 8$. We can then verify these cases directly using Mathematica to check these cases and we find no obstructions.

[27, ∞). We will apply Remark 2.2. As

$$\sqrt{27} \geq \frac{7.5}{\sqrt{13}} \left(2 + \frac{7.5}{d} \right) \quad \text{for } d \geq 18,$$

Remark 2.2 implies we only need to verify $N_k(1, a^2) \leq (a/13)M_k(1, 6.5)$ for all

$$k \leq \frac{18^2}{26} + \frac{7.5 \cdot 18}{13} + \frac{6.5^2 - 13 + 1}{26} < 25.$$

For $a^2 \geq 27$ and $k \leq 25$, we have

$$N_k(1, a^2) = k \leq \sqrt{\frac{27}{13}} M_k(1, 6.5) \leq \frac{a}{\sqrt{13}} M_k(1, 6.5).$$

This completes the proof that $d(a, b) = \sqrt{a/13}$ for $a \in [27, \infty)$.

6. Conjectures

We now present some conjectures concerning exactly when an ellipsoid embeds into a polydisc.

Extensions of Theorem 1.1. To consider an interesting refinement of Theorem 1.1, define

$$V(b) = \inf \left\{ A : d(a, b) = \sqrt{\frac{a}{2b}} \quad \text{for } a \geq A \right\}.$$

Theorem 1.1 implies $V(b) \leq \frac{9}{2}(b + 2 + 1/b)$.

Proposition 6.1. For $b \geq 1$,

$$V(b) \geq 2b \left(\frac{2[b] + 2[\sqrt{2b} + \{b\}] - 1}{b + [b] + [\sqrt{2b} + \{b\}] - 1} \right)^2.$$

Proof.

$$\begin{aligned} d(2[b] + 2[\sqrt{2b} + \{b\}] - 1, b) &\geq \frac{N_{2[b] + 2[\sqrt{2b} + \{b\}] - 1}(1, 2[b] + 2[\sqrt{2b} + \{b\}] - 1)}{M_{2[b] + 2[\sqrt{2b} + \{b\}] - 1}(1, b)} \\ &= \frac{2[b] + 2[\sqrt{2b} + \{b\}] - 1}{b + [b] + [\sqrt{2b} + \{b\}] - 1} \\ &> \sqrt{\frac{2[b] + 2[\sqrt{2b} + \{b\}] - 1}{2b}}. \end{aligned}$$

This implies

$$V(b) \geq 2b \left(\frac{2[b] + 2[\sqrt{2b} + \{b\}] - 1}{b + [b] + [\sqrt{2b} + \{b\}] - 1} \right)^2. \quad \square$$

Experimental evidence seems to suggest that for $b > 1$ this bound is sharp.

Conjecture 6.2. For $b > 1$,

$$V(b) = 2b \left(\frac{2\lfloor b \rfloor + 2\lceil \sqrt{2b} + \{b\} \rceil - 1}{b + \lfloor b \rfloor + \lceil \sqrt{2b} + \{b\} \rceil - 1} \right)^2.$$

Generalizations of Theorem 1.2. The methods used to compute the graph of $d(a, 6.5)$ should extend for the most part to any b . In light of those techniques, experimental evidence, and a conjecture regarding $d(a, b)$ for $b \in \mathbb{Z}$ by David Frenkel and Felix Schlenk relayed to us by Daniel Cristofaro-Gardiner, we offer a conjecture regarding the graph of $d(a, b)$ for $b \geq 6$; see Figure 2.

Conjecture 6.3. For $b \geq 6$, we have $d(a, b) = \sqrt{a/(2b)}$ with the exception that

$$d(a, b) = 1 \quad \text{for } a \in [1, b + \lfloor b \rfloor].$$

For $k \in \mathbb{Z}$, with $0 \leq k < \sqrt{2b} + \{b\}$, we have

$$d(a, b) = \frac{a}{b + \lfloor b \rfloor + k} \quad \text{for } a \in [\alpha_k, 2(\lfloor b \rfloor + k) + 1],$$

$$d(a, b) = \frac{2(\lfloor b \rfloor + k) + 1}{b + \lfloor b \rfloor + k} \quad \text{for } a \in [2(\lfloor b \rfloor + k) + 1, \beta_k],$$

where

$$\alpha_0 = b + \lfloor b \rfloor, \quad \alpha_1 = \beta_0 = \frac{(b + \lfloor b \rfloor + 1)(2\lfloor b \rfloor + 1)}{b + \lfloor b \rfloor},$$

$$\alpha_k = \frac{(b + \lfloor b \rfloor + k)^2}{2b} \quad \text{for } k \geq 2, \quad \beta_k = 2b \left(\frac{2(\lfloor b \rfloor + k) + 1}{b + \lfloor b \rfloor + k} \right)^2 \quad \text{for } k \geq 1.$$

For integers m , if

$$b \in \left[m - \frac{m}{(m+1)^2}, m + \frac{1}{2+m} \right],$$

let $b = m + \varepsilon$. Then

$$d(a, b) = \frac{ma + 1}{2m^2 + (2 + \varepsilon)m + \varepsilon} \quad \text{for } a \in [\alpha^*, 2m + 4],$$

$$d(a, b) = \frac{m(2m + 4) + 1}{2m^2 + (2 + \varepsilon)m + \varepsilon} \quad \text{for } a \in [2m + 4, \beta^*],$$

where

$$\alpha^* = \frac{1}{2(2m^3 + 2m^2\varepsilon)} \left(8m^3 + 4m^2 + 8m^2\varepsilon + 4m^3\varepsilon + \varepsilon^2 + 2m\varepsilon^2 + b^2\varepsilon^2 - (1+m)(2m+\varepsilon) \right. \\ \left. \times \sqrt{-4m^2 + 8m^3 + 4m^4 - 4m\varepsilon + 8m^2\varepsilon + 4m^3\varepsilon + \varepsilon^2 + 2m\varepsilon^2 + m^2\varepsilon^2} \right),$$

$$\beta^* = \frac{2(\varepsilon + m + 8m\varepsilon + 8m^2 + 20m^2\varepsilon + 16m^3\varepsilon + 16m^4 + 4m^4\varepsilon + 4m^5)}{(1+m)^2(2m+\varepsilon)^2}.$$

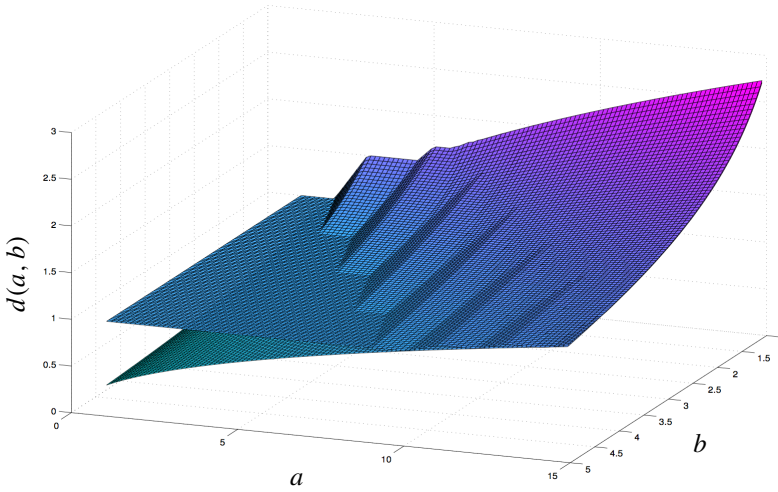


Figure 2. Approximate plot of the graph of $d(a, b)$.

We note that Conjecture 6.3 implies Conjecture 6.2 for $b \geq 6$. Furthermore, we prove that the conjecture is a lower bound for $d(a, b)$.

Proposition 6.4. For $b \geq 6$, we have $d(a, b) \geq \sqrt{a/(2b)}$ and

$$d(a, b) \geq 1 \quad \text{for } a \in [1, b + \lfloor b \rfloor].$$

For $k \in \mathbb{Z}$, with $0 \leq k < \sqrt{2b} + \{b\}$, we have

$$d(a, b) \geq \frac{a}{b + \lfloor b \rfloor + k} \quad \text{for } a \in [\alpha_k, 2(\lfloor b \rfloor + k) + 1],$$

$$d(a, b) \geq \frac{2(\lfloor b \rfloor + k) + 1}{b + \lfloor b \rfloor + k} \quad \text{for } a \in [2(\lfloor b \rfloor + k) + 1, \beta_k],$$

where $\alpha_k, \beta_k, \alpha^*, \beta^*$ are as in Conjecture 6.3. For integers m , if

$$b \in \left[m - \frac{m}{(m+1)^2}, m + \frac{1}{2+m} \right],$$

let $b = m + \varepsilon$. Then

$$d(a, b) \geq \frac{ma + 1}{2m^2 + (2 + \varepsilon)m + \varepsilon} \quad \text{for } a \in [\alpha^*, 2m + 4],$$

$$d(a, b) \geq \frac{m(2m + 4) + 1}{2m^2 + (2 + \varepsilon)m + \varepsilon} \quad \text{for } a \in [2m + 4, \beta^*].$$

Proof. We know that $d(a, b) \geq \sqrt{a/(2b)}$ because symplectic embeddings are volume preserving. We also have

$$d(a, b) \geq \frac{N_1(1, a)}{M_1(a, b)} = \frac{1}{1} = 1.$$

Additionally, for $k \in \mathbb{Z}$, with $\lfloor k \rfloor < \sqrt{2b} + \{b\}$, and $a \in [2(\lfloor b \rfloor + k), 2(\lfloor b \rfloor + k) + 1]$, we have

$$d(a, b) \geq \frac{N_{2(\lfloor b \rfloor + k) + 1}(1, a)}{M_{2(\lfloor b \rfloor + k) + 1}(1, b)} = \frac{a}{b + \lfloor b \rfloor + k}.$$

Thus,

$$d(a, b) \geq 1 \quad \text{for } a \in [b + \lfloor b \rfloor, 2\lfloor b \rfloor + 1], \quad k = 0,$$

$$d(a, b) \geq \frac{2\lfloor b \rfloor + 1}{b + \lfloor b \rfloor} \quad \text{for } a \in \left[\frac{(b + \lfloor b \rfloor + 1)(2\lfloor b \rfloor + 1)}{b + \lfloor b \rfloor}, 2\lfloor b \rfloor + 3 \right], \quad k = 1,$$

$$d(a, b) \geq \sqrt{\frac{a}{2b}} \quad \text{for } a \in [\alpha_k, 2(\lfloor b \rfloor + k) + 1], \quad k \geq 2.$$

We also have, for $a \in [2(\lfloor b \rfloor + k) + 1, \infty)$,

$$d(a, b) \geq \frac{N_{2(\lfloor b \rfloor + k) + 1}(1, a)}{M_{2(\lfloor b \rfloor + k) + 1}(1, b)} = \frac{2(\lfloor b \rfloor + k) + 1}{b + \lfloor b \rfloor + k}.$$

Thus,

$$d(a, b) \geq \sqrt{\frac{a}{2b}} \quad \text{for } a \in [2(\lfloor b \rfloor + k) + 1, \beta_k].$$

Furthermore, if

$$b \in \left[m - \frac{m}{(m+1)^2}, m + \frac{1}{2+m} \right]$$

for some $m \in \mathbb{Z}$ and $a \in [2m + 4 - 1/m, 2m + 4]$, then

$$d(a, b) \geq \frac{N_{(m+1)^3}(1, a)}{M_{(m+1)^3}(1, b)} = \frac{ma + 1}{2m^2 + (2 + \varepsilon)m + \varepsilon},$$

so

$$d(a, b) \geq \sqrt{\frac{a}{2b}} \quad \text{for } a \in [\alpha^*, 2m + 4].$$

We also have, for $a \in [2m + 4m\beta^*]$,

$$d(a, b) \geq \frac{N_{(m+1)^3}(1, a)}{M_{(m+1)^3}(1, b)} = \frac{m(2m + 4) + 1}{2m^2 + (2 + \varepsilon)m + \varepsilon}.$$

Thus,

$$d(a, b) \geq \sqrt{\frac{a}{2b}} \quad \text{for } a \in [2m + 4, \beta^*]. \quad \square$$

Acknowledgments

We wish to thank Daniel Cristofaro-Gardiner for his helpful explanations, reference suggestions, encouragement and patience. We also thank the NSF, Michael Hutchings and UC Berkeley for providing the opportunity to work on symplectic embedding problems this summer.

References

- [Beck and Robins 2007] M. Beck and S. Robins, *Computing the continuous discretely: Integer-point enumeration in polyhedra*, Springer, New York, 2007. MR Zbl
- [Buse and Hind 2013] O. Buse and R. Hind, “Ellipsoid embeddings and symplectic packing stability”, *Compos. Math.* **149**:5 (2013), 889–902. MR Zbl
- [Cristofaro-Gardiner and Kleinman 2013] D. Cristofaro-Gardiner and A. Kleinman, “Ehrhart polynomials and symplectic embeddings of ellipsoids”, preprint, 2013. arXiv
- [Frenkel and Müller 2012] D. Frenkel and D. Müller, “Symplectic embeddings of 4-dimensional ellipsoids into cubes”, preprint, 2012. arXiv
- [Hutchings 2014] M. Hutchings, “Lecture notes on embedded contact homology”, pp. 389–484 in *Contact and symplectic topology*, edited by F. Bourgeois et al., Bolyai Soc. Math. Stud. **26**, János Bolyai Math. Soc., Budapest, 2014. MR Zbl
- [Li and Li 2002] B.-H. Li and T.-J. Li, “Symplectic genus, minimal genus and diffeomorphisms”, *Asian J. Math.* **6**:1 (2002), 123–144. MR Zbl
- [McDuff 2011] D. McDuff, “The Hofer conjecture on embedding symplectic ellipsoids”, *J. Differential Geom.* **88**:3 (2011), 519–532. MR Zbl
- [McDuff and Schlenk 2012] D. McDuff and F. Schlenk, “The embedding capacity of 4-dimensional symplectic ellipsoids”, *Ann. of Math. (2)* **175**:3 (2012), 1191–1282. MR Zbl

Received: 2014-09-08 Revised: 2015-06-21 Accepted: 2015-07-01

burkham2@uw.edu

*Mathematics Department, University of Washington,
Padelford Hall C-110, Seattle, WA 98195, United States*

University of California, Santa Cruz, CA 95064, United States

mtimmons@mit.edu

*Massachusetts Institute of Technology, 450 Memorial Drive,
Cambridge, MA 02139, United States*

Characterizations of the round two-dimensional sphere in terms of closed geodesics

Lee Kennard and Jordan Rainone

(Communicated by Kenneth S. Berenhaut)

The question of whether a closed Riemannian manifold has infinitely many geometrically distinct closed geodesics has a long history. Though unsolved in general, it is well understood in the case of surfaces. For surfaces of revolution diffeomorphic to the sphere, a refinement of this problem was introduced by Borzellino, Jordan-Squire, Petrics, and Sullivan. In this article, we quantify their result by counting distinct geodesics of bounded length. In addition, we reframe these results to obtain a couple of characterizations of the round two-sphere.

Introduction

All closed Riemannian manifolds contain a closed geodesic. If the manifold is not simply connected, any length-minimizing representative of a nontrivial homotopy class is a closed geodesic. In the simply connected case, this is already a nontrivial result.

A more difficult question is whether there exist infinitely many closed geodesics. To avoid over-counting, one considers two geodesics *geometrically distinct* if their images are distinct. This brings us to the well-known question of whether there exist infinitely many geometrically distinct closed geodesics. In this article, we restrict our attention to surfaces, but we refer the reader to [Oancea 2015, Chapter 2] for a survey and a guide to the literature on the problem.

For surfaces with genus $g \geq 1$, one uses the infinitude of the fundamental group and a length-minimization argument to construct infinitely many geometrically distinct closed geodesics. For the torus, it follows that the number of such geodesics of length at most ℓ grows quadratically in ℓ (see [Berger 2010, Chapter XII.5.A]). For $g \geq 2$, Katok proved that this number actually grows exponentially in ℓ (see Remark 0.3 below).

In the remaining case, when the surface is the sphere, this question was answered affirmatively by Bangert [1993] and Franks [1992] (cf. [Berger 2010; Hingston

MSC2010: 53C20, 58E10.

Keywords: closed geodesics, surface of revolution.

1993a]). Hingston [1993b] then proved a quantified version of this result: given any metric on \mathbb{S}^2 , the number of geometrically distinct closed geodesics of length at most ℓ is asymptotically at least $c\ell/\log \ell$ for some constant $c > 0$.

In this article, we consider refinements of these results. As motivation, consider a surface of revolution. Each profile curve connecting the poles extends to a closed geodesic. In particular, the results of Bangert, Franks and Hingston are trivial in this setting. On the other hand, all of these geodesics are in some sense the same. This motivates the following definition: for a closed Riemannian manifold M , we say that two geodesics on M are *strongly geometrically distinct* if there is no isometry taking the image of one to the image of the other.

For metrics with finite isometry group, one has immediate analogues of the results above. For metrics with infinite symmetry, it is unclear whether there exist infinitely many strongly geometrically distinct geodesics. For example, the constant curvature metric on \mathbb{S}^2 has only one closed geodesic in this sense. Borzellino et al. [2007] proved that all surfaces of revolution diffeomorphic to \mathbb{S}^2 , except for the round spheres, have infinitely many strongly geometrically distinct geodesics. Our main result is a quantification of this result, as well as a straightforward observation that it extends to all closed, orientable surfaces with continuous (equivalently infinite) symmetry.

Main Theorem. *Let M be an orientable, compact surface with infinite isometry group. Let $N(\ell)$ denote the number of strongly geometrically distinct closed geodesics on M of length less than or equal to ℓ . One of the following occurs:*

- (1) *M is isometric to a round sphere, and $N(\ell) = 1$ for all sufficiently large $\ell > 0$.*
- (2) *There is a constant $c > 0$ such that $N(\ell) \geq c\ell^2$ for all sufficiently large $\ell > 0$.*

We make a few remarks.

Remark 0.1. In the nonorientable case, one applies the theorem to the orientable double cover to obtain an analogous characterization of the real projective plane with constant curvature.

Remark 0.2. It is well known that a closed, orientable surface M can have infinite isometry group only if M is diffeomorphic to \mathbb{S}^2 or the torus T^2 (see Lemma 1.1). In the latter case, a simple extension of a standard argument shows the Main Theorem holds. However the argument we provide for \mathbb{S}^2 carries over with little effort to the case of T^2 , so we include it in Section 3 for completeness.

Remark 0.3. For a compact surface M with genus $g \geq 2$, the isometry group is finite, so $N(\ell)$ is related to the number $n(\ell)$ of geometrically distinct closed geodesics on M of length at most ℓ by the relation

$$N(\ell) \leq n(\ell) \leq CN(\ell),$$

where C denotes the number of elements in the isometry group. Hence asymptotics on $n(\ell)$ imply asymptotics on $N(\ell)$, up to multiplicative constant. For a metric on M with constant curvature -1 , Margulis showed that the function $n(\ell)$ is asymptotic to ce^ℓ/ℓ for some constant c ; that is, $n(\ell)/(ce^\ell/\ell) \rightarrow 1$ as $\ell \rightarrow \infty$ (see [Margulis 1969]; cf. [Katok 1988, Section 1]). In particular, $n(\ell) \leq e^\ell$ for all sufficiently large ℓ . On the other hand, Katok [1982] showed that, for any metric on M with the same area as the constant curvature -1 metric,

$$\liminf_{\ell \rightarrow \infty} \log(n(\ell))/\ell \geq 1,$$

with equality if and only if the metric has constant curvature -1 (cf. [Berger 2010, Chapter XII.5.B]). As a consequence, for the case of nonconstant curvature, there exists a constant $a > 1$ such that $n(\ell) \geq e^{a\ell}$ for all sufficiently large ℓ . Hence for both \mathbb{S}^2 and surfaces of genus $g \geq 2$, there is a sense in which the constant curvature metric is characterized by having the fewest closed geodesics. We do not know whether the constant curvature metrics on T^2 have a similar characterization.

Consider now a metric on \mathbb{S}^2 with infinite isometry group. The metric takes the form $ds^2 + h(s)^2 d\theta^2$ and one can check that the arguments in [Borzellino et al. 2007] for a surface of revolution carry over to this slightly more general case to show that infinitely many strongly geometrically distinct closed geodesics exist, i.e., $\lim_{\ell \rightarrow \infty} N(\ell) = \infty$. In Section 2, we summarize their argument and supplement it where needed to prove the claimed lower bound on the growth rate of $N(\ell)$.

Before starting the proof, we point out that this theorem, combined with the work of Hingston and Katok, immediately implies the following:

Corollary. *Let M be an orientable, compact surface. Either M is isometric to a round sphere and $N(\ell) = 1$ for all sufficiently large $\ell > 0$, or there exists a constant $c > 0$ such that $N(\ell) \geq c\ell/\log \ell$ for all sufficiently large $\ell > 0$.*

1. Preliminaries on Lie group actions

In this section, we gather some results on isometric actions by Lie groups that are required for the proofs. We summarize the results here:

Lemma 1.1. *If M is a closed, orientable Riemannian manifold of dimension two with infinite isometry group G , then the identity component $G_0 \subseteq G$ contains a circle \mathbb{S}^1 , and one of the following occurs:*

- (1) M is isometric to a round \mathbb{S}^2 and $\dim G = 3$.
- (2) M is diffeomorphic to \mathbb{S}^2 but not isometric to a round \mathbb{S}^2 , $\dim G = 1$, and the fixed-point set of \mathbb{S}^1 is a pair of isolated points.
- (3) M is diffeomorphic to a torus, and the fixed-point set of \mathbb{S}^1 is empty.

In particular, M cannot have genus $g \geq 2$.

To prove this lemma, suppose M is a closed Riemannian manifold of dimension two with infinite isometry group G . A theorem of Myers and Steenrod states that G is a compact Lie group (see [Kobayashi 1972, Chapter II, Section 1]). Let $G_0 \subseteq G$ denote the identity component. By compactness, G has only finitely many components. Since G is infinite, this implies G_0 has positive dimension. In particular, the maximal torus theorem implies G_0 contains a circle \mathbb{S}^1 .

This circle acts isometrically on M , and its fixed-point set

$$F = \{p \in M \mid e^{it}(p) = p \text{ for all } e^{it} \in \mathbb{S}^1\}$$

equals the zero set of the associated Killing field X on M defined by

$$X(p) = \left. \frac{d}{dt} \right|_{t=0} (e^{it}(p)).$$

Moreover, F consists of isolated points, and the number of these points equals the Euler characteristic of M (see [Kobayashi 1972, Chapter II, Theorems 5.3 and 5.5]). Since the Euler characteristic of M equals $2 - 2g$, where g is the genus, it follows either that M is diffeomorphic to \mathbb{S}^2 and F is a pair of isolated points or that M is diffeomorphic to T^2 and F is empty.

It suffices to show that $\dim G = 3$ if and only if M is a round \mathbb{S}^2 , and that $\dim G = 2$ only if M is diffeomorphic to T^2 . Regarding the first of these claims, we note that a round \mathbb{S}^2 has isometry group $O(3)$, which is three-dimensional. Conversely, it is a classical fact that if the isometry group of a compact two-manifold is three-dimensional, then M is either \mathbb{S}^2 or the real projective plane $\mathbb{R}\mathbb{P}^2$ equipped with a metric of constant curvature (see [Kobayashi 1972, Chapter II, Theorem 3.1]). If, moreover, M is orientable, as in Lemma 1.1, then we conclude that M is isometric to a round \mathbb{S}^2 .

Suppose now that $\dim G = 2$. The only compact, connected, two-dimensional Lie group is the two-torus, so $G_0 = T^2$ (see [Bröcker and tom Dieck 1985, page 169]). Since G_0 acts effectively on M and has the same dimension as M , it follows that G_0 acts transitively on M and hence that the Gauss curvature is constant. By the Gauss–Bonnet theorem and the fact that the genus $g \leq 1$, either M is a round \mathbb{S}^2 or a flat T^2 . In the first of these cases, we have $\dim G = 3$, a contradiction to the assumption that $\dim G = 2$. Hence M is isometric to a torus with constant zero curvature.

2. Proof of the Main Theorem for the sphere

Assume that M is a Riemannian manifold diffeomorphic to \mathbb{S}^2 with infinite isometry group. Let $\{p, q\} \subseteq M$ denote the fixed point set of this circle action according to Lemma 1.1. Choose a minimal geodesic c from p to q . By rescaling the metric if necessary, assume that c is defined on $[0, \pi]$ and that $c(0) = p$ and $c(\pi) = q$.

There exists a smooth function $h : (0, \pi) \rightarrow (0, \infty)$ and an isometric covering map

$$\begin{aligned} \sigma : ((0, \pi) \times \mathbb{R}, ds^2 + h(s)^2 d\theta^2) &\rightarrow M \setminus \{p, q\}, \\ (s, \theta) &\mapsto e^{i\theta} \cdot c(s), \end{aligned}$$

where the dot denotes the action of the circle element $e^{i\theta}$ on $c(s)$. Since M is smooth at $p = c(0)$ and $q = c(\pi)$, we conclude that the extended function $h : [0, \pi] \rightarrow \mathbb{R}$ satisfies $h(0) = h(\pi) = 0$ and $h'(0) = -h'(\pi) = 1$ (see [Petersen 2016, Section 1.4.4]). The strategy now is to follow the proof in [Borzellino et al. 2007], which covers the case of a surface of revolution. Note that, for a surface of revolution, $h(s)$ represents one coordinate of a unit-speed curve in the plane and hence satisfies the condition that $|h'(s)| \leq 1$ (see [Petersen 2016, Section 1.4.4]). Although we are considering a more general class of surfaces, the arguments of [Borzellino et al. 2007] extend to our situation. We summarize the proof here since our strategy is simply to supplement it, as needed, in order to prove the Main Theorem.

In the coordinates induced by σ , the geodesic equations are

$$\begin{aligned} s''(t) &= h(s(t))h'(s(t))\theta'(t)^2, \\ \theta''(t) &= -2\frac{h'(s(t))}{h(s(t))}s'(t)\theta'(t). \end{aligned}$$

The meridians, $\gamma(t) = \sigma(t, \theta_0)$, satisfy these equations and extend to closed geodesics passing through both poles, p and q . Since θ_0 is arbitrary, we have by uniqueness that meridians are the only geodesics that pass through the poles. In the rest of this section, we consider those geodesics that do not pass through the poles. Since σ defines an isometric covering map onto $M \setminus \{p, q\}$, we can write a geodesic $\gamma(t)$ as $\sigma(s(t), \theta(t))$ for smooth functions $s : \mathbb{R} \rightarrow (0, \pi)$ and $\theta : \mathbb{R} \rightarrow \mathbb{R}$. For example, the parallels given by $\gamma(t) = \sigma(s_0, t/h(s_0))$ are closed geodesics provided that $h'(s_0) = 0$. Another example of a geodesic is provided in Figure 1.

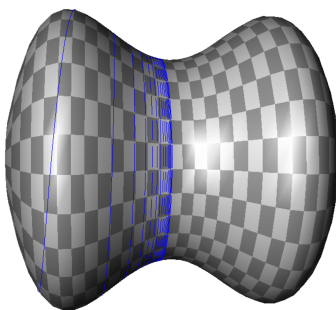


Figure 1. A geodesic asymptotic to a parallel. The surface is \mathbb{S}^2 equipped with a rotationally symmetric metric.

An important consequence of the geodesic equations is Clairaut’s relation. This states that, for each nonmeridian geodesic γ , there exists a constant $c_\gamma > 0$ such that

$$h(s(t)) \cos \alpha(t) = c_\gamma,$$

where $\alpha(t)$ is the angle between $\gamma'(t)$ and the coordinate vector field σ_θ at $\gamma(t)$. Since the cosine function is bounded, $h(s(t))$ cannot go to zero; hence any nonmeridian curve has its s -coordinate bounded by some interval

$$[s_0(\gamma), s_1(\gamma)] = [\inf s(t), \sup s(t)] \subseteq (0, \pi).$$

Further analysis shows the following.

Lemma 2.1 (Clairaut). *For $a \in (0, \pi)$, let γ_a be a unit-speed geodesic starting with s -coordinate a and initial direction $\gamma'(0)$ in the θ -direction. One of the following occurs:*

- (1) **parallel:** $h'(a) = 0$, and $s(t) = a$ for all t .
- (2) **asymptotic:** $h'(a) > 0$ (resp. < 0) and there exists $b = b(a) > a$ (resp. $< a$) such that $h'(b) = 0$ and $s(t) \rightarrow b$ as $t \rightarrow \infty$.
- (3) **oscillating:** $h'(a) > 0$ (resp. < 0) and there exists $b = b(a) > a$ (resp. $< a$) such that $h'(b) < 0$ (resp. > 0) and $s(t)$ oscillates between a and b , achieving these extremal values at integral multiples of some time, denoted $T(a)$.

According to this result, we refer to the parameter $a \in (0, \pi)$ as parallel, asymptotic, or oscillating. Following [Borzellino et al. 2007, Proposition 3.1], we let $U \subseteq (0, \pi)$ denote the subset consisting of oscillating $a \in (0, \pi)$ for which $h'(a) > 0$ and $h'(b(a)) < 0$, where $b(a) = \inf\{b > a \mid h(b) = h(a)\}$. Geometrically, the s -coordinate of γ_a oscillates between a and $b(a)$. It follows that $U \subseteq (0, \pi)$ is an open set and that the function $a \mapsto b(a)$ on U is smooth. Indeed, this function is given by h composed with a local inverse of h , and so it is smooth by the inverse function theorem. Figure 2 indicates the region U for a function $h(s)$ corresponding to the dumbbell shape from Figure 1.

For each $a \in U$, let $\gamma_a(t) = \sigma(s(t), \theta(t))$ be as in Lemma 2.1 and define

$$R(a) = 2 \int_0^{T(a)} \theta'(t) dt \quad \text{and} \quad L(a) = 2T(a) = 2 \int_0^{T(a)} 1 dt,$$

where $T(a)$ is the time referred to in the third conclusion of Lemma 2.1. This defines two functions $R : U \rightarrow \mathbb{R}$ and $L : U \rightarrow \mathbb{R}$. The geometric interpretation of these functions is as follows. The quantity $2T(a)$ denotes the time required for a geodesic starting at $s = a$ and parallel to σ_θ to have its s -coordinate go to $b(a)$ and back to a . We call this a “full trip”. It then follows by symmetry that $R(a)$ and $L(a)$ denote the total rotation and length of the geodesic on a full trip. In [Borzellino

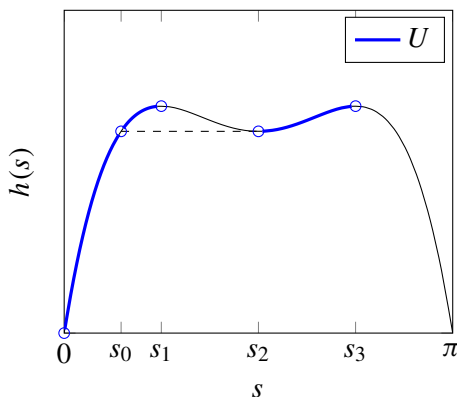


Figure 2. Example of function $h(s)$ corresponding to a surface of revolution with the shape of a dumbbell, as in Figure 1. Here, s is the arclength coordinate. The value $a = s_0$ corresponds to an asymptotic geodesic as in Lemma 2.1, and the values $a \in \{s_1, s_2, s_3\}$ correspond to parallel geodesics. The blue region is U , the set of oscillating values of a for which $h'(a) > 0$.

et al. 2007], the authors prove that $R(a)$ is a continuous function of a . For our purposes, we also need that $L(a)$ is continuous.

Lemma 2.2. *The functions $L, R : U \rightarrow \mathbb{R}$ are continuous.*

Proof. The proofs for R and L are similar, so we only prove it for L . Fix $a \in U$. Choose a nontrivial interval $[a_1, a_2] \subseteq U$ containing a on which $h' \geq c_1 > 0$. We prove now that L is continuous on $[a_1, a_2]$.

To do this, we rewrite expression for $L(a)$. First, the unit-speed condition implies $1 = |\gamma'_a(t)|^2 = s'(t)^2 + h(s(t))^2\theta'(t)^2$. Since $s(t)$ is increasing from $t = 0$ to $t = T(a)$, this implies

$$s'(t) = \sqrt{1 - h(s(t))^2\theta'(t)^2}.$$

Next, the second geodesic equation implies $\frac{d}{dt}(h(s(t))^2\theta'(t)) = 0$. As a result, $h(s(t))^2\theta'(t)$ equals a constant C . At $t = 0$, the unit-speed condition implies $\theta'(0) = 1/h(s(0)) = 1/h(a)$, so we have $C = h(a)$. Putting this together, we obtain

$$s'(t) = \sqrt{1 - h(a)^2/h(s(t))^2}.$$

Finally, we use this expression in order to apply the change of variables $s = s(t)$ to the integral $L = 2 \int_0^{T(a)} dt$. This gives us the expression

$$L = 2 \int_a^{b(a)} \frac{ds}{\sqrt{1 - h(a)^2/h(s)^2}}.$$

Regarding the right side as a function of a , we may write $L(a) = 2 \int_a^{b(a)} l(a, s) ds$, where $l(a, s)$ is given by $h(s)/\sqrt{h(s)^2 - h(a)^2}$. This integral is improper at both endpoints, so we proceed by proving the following two claims:

- (1) For all sufficiently small $\delta > 0$, the integral $L_\delta(a) = 2 \int_{a+\delta}^{b(a)-\delta} l(a, s) ds$ is smooth.
- (2) The functions L_δ converge uniformly to L on $[a_1, a_2]$.

The first claim follows from the Leibniz integral rule since $l(a, s)$ is a smooth function on the set $\{(a, s) | a \in [a_1, a_2], a + \delta \leq s \leq b(a) - \delta\}$. To prove the second claim, it suffices to prove that $\int_a^{a+\delta} l(a, s) ds \rightarrow 0$ and $\int_{b(a)-\delta}^{b(a)} l(a, s) ds \rightarrow 0$ uniformly in $a \in [a_1, a_2]$ as δ goes to 0. These claims are proven similarly, so we only prove the first. The second only requires the additional fact that $b(a)$ depends smoothly on a .

Observe that $l(a, s)$ is nonnegative and bounded above as

$$l(a, s) = \frac{h(s)}{\sqrt{h(s)^2 - h(a)^2}} \leq \frac{1}{2c_1} \frac{2h(s)h'(s)}{\sqrt{h(s)^2 - h(a)^2}}.$$

Integrating this expression and applying the change of variables $y = h(s)^2 - h(a)^2$, we conclude that

$$\int_a^{a+\delta} l(a, s) ds \leq \frac{1}{2c_1} \int_0^{h(a+\delta)^2 - h(a)^2} \frac{dy}{\sqrt{y}} = \frac{\sqrt{h(a+\delta)^2 - h(a)^2}}{c_1}.$$

Since h is smooth and hence uniformly continuous on $[0, \pi]$, this last quantity converges to 0 uniformly in a as $\delta \rightarrow 0$. □

We proceed to the proof of the Main Theorem, that the number $N(\ell)$ of strongly geometrically distinct closed geodesics grows quadratically in ℓ . The idea is to show, for all large $\ell > 0$, that a large number of values of a exist such that $a \in U$, $R(a) = 2\pi(p/q)$ for some rational p/q , and $L(a) \leq \ell/q$. These three conditions imply that any choice of γ_a as in Lemma 2.1 is oscillating, closes up after q full trips, and is a closed geodesic with length at most ℓ .

First, we dispose of the case where the isometry group G satisfies $\dim G \neq 1$. By Lemma 1.1, we have $\dim G = 3$ and that M is a round sphere. In this case, the isometry group is $O(3)$ or $SO(3)$, and every unit-speed geodesic can be carried to any other by an isometry, so $N(\ell) = 1$ for all ℓ larger than $2\pi r$, where $1/r^2$ is the Gauss curvature of M . This completes the proof of the Main Theorem in this case.

We assume from now on that $\dim G = 1$. As a result, the identity component $G_0 \subseteq G$ equals the circle group. By compactness, G has only finitely many components. In particular, for each oscillating value of a as above, at most finitely many other such values result in geodesics that are not strongly geometrically distinct from γ_a . This issue results in a multiplicative factor (equal to the number

of components in the isometry group) in our estimates. Since the Main Theorem involves an unknown multiplicative constant anyway, we simply assume, without loss of generality, that the isometry group equals the circle.

The proof is carried out in three cases, which are based roughly on the setup in [Borzellino et al. 2007]. One key step is to prove that there exists an asymptotic geodesic if h has more than one critical point. This actually need not be the case. Indeed, a capped cylinder provides a counterexample, since every critical point is a local maximum and hence not a limiting value of an asymptotic geodesic. This problem is easy to fix, however, by breaking the proof into cases as follows.

Lemma 2.3. *If h has infinitely many critical points, then $N(\ell) = \infty$ for all sufficiently large $\ell > 0$.*

Proof. If $h'(a) = 0$, then $\gamma_a(t) = \sigma(a, t/h(a))$ is a closed geodesic of length $2\pi h(a)$. Moreover, the image of γ_a maps to itself under any isometry, so distinct values of a yield strongly geometrically distinct closed geodesics. The result follows since h is bounded on $[0, \pi]$. \square

Lemma 2.4. *If h has finitely many critical points, and R is locally constant, then $N(\ell) = \infty$ for all sufficiently large $\ell > 0$.*

Proof. In this case, the argument in [Borzellino et al. 2007, Corollaries 4.4 and 4.5] is valid since the critical points are isolated. Indeed, first suppose that h has more than one critical point (as in Figure 2). The arguments there show that M has an asymptotic geodesic and hence that R is unbounded on U . However, Lemma 2.1 and the assumptions of this lemma imply that R takes on only finitely many values, so this is a contradiction. Assume instead that h has a unique critical point, s_0 (as in Figure 3 below). It follows as in [Borzellino et al. 2007, Corollary 5.4] that $U = (0, s_0)$ and that $R(a) = \lim_{a' \rightarrow 0} R(a') = 2\pi$ for all $a \in (0, s_0)$. But L is continuous on $(0, s_0)$ and hence on $[s_0/3, s_0/2]$, so there exist infinitely many strongly geometrically distinct closed geodesics of length at most L_0 , where $L_0 = \max\{L(s) \mid s \in [s_0/3, s_0/2]\} < \infty$. \square

Lemma 2.5. *If h has finitely many critical points and R is not locally constant, then there exists a constant $c > 0$ such that $N(\ell) \geq c\ell^2$ for all sufficiently large $\ell > 0$.*

Proof. Choose a closed interval $I' \subseteq U$ that is mapped by R to some nontrivial interval $I \subseteq \mathbb{R}$. Let $2\pi(p/q) \in I$. Each $a \in U$ that is mapped by R to $2\pi(p/q)$ corresponds to a closed geodesic of length $qL(a)$. Since L is continuous on I' , this length is at most qL_0 , where L_0 is the maximum value of L on I' . This length is at most ℓ if and only if $q \leq \lfloor \ell/L_0 \rfloor$. To estimate $N(\ell)$ from below, it suffices to count the number of rationals $p/q \in 1/(2\pi)I$ with $q \leq \lfloor \ell/L_0 \rfloor$. By Lemma 2.6 below, there is a constant c' such that the number of such rationals is at least $c'(\lfloor \ell/L_0 \rfloor)^2$

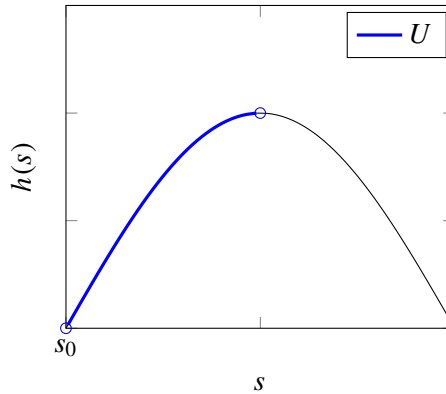


Figure 3. An example of a profile curve $h(s)$ with a unique critical point. As in Figure 2, s is the arclength parameter and U is the set of oscillating s -values a for which $h'(a) > 0$.

for all sufficiently large ℓ . Taking $c = \frac{1}{2}c'/L_0^2$, we conclude that $N(\ell) \geq c\ell^2$ for all sufficiently large $\ell > 0$. □

As indicated in the previous proof, it suffices to prove the following counting lemma.

Lemma 2.6. *Inside any connected, nontrivial interval $I \subseteq \mathbb{R}$, there exist constants $c > 0$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, there are at least cn^2 rational numbers in I with denominator at most n .*

Proof. The proof uses Farey fractions. Let F_n denote the set of rationals a/b written in reduced form such that $0 \leq a \leq b \leq n$. It is easy to see that the number of elements in F_n satisfies

$$|F_n| = 1 + \sum_{k=1}^n \phi(k),$$

where $\phi(k)$ is the Euler totient function, given by the number of integers $1 \leq i \leq k$ coprime to k . According to Walfisz [1963],

$$\sum_{k=1}^n \phi(k) = \frac{3}{\pi^2}n^2 + O(n(\log n)^{2/3}(\log \log n)^{4/3}).$$

In particular, it follows that constants $c_1 > 0$ and $n_0 > 0$ exist such that $|F_n| > c_1n^2$ for all $n \geq n_0$.

The idea now is to inject F_n into I in a controlled way. First, it is clear that the conclusion of the lemma holds for I if and only if it holds for $\{1+i \mid i \in I\}$. Hence, we assume without loss of generality that $I \not\subseteq (-\infty, 0]$. Choose positive integers a

and b such that I contains the interval $[a/b, (a+1)/b]$. Set $c = \frac{1}{2}(c_1/b^2)$, and choose $n_0 \geq n_1$ such that $\lfloor n/b \rfloor \geq n_1$ and $c_1(n/b - 1)^2 > cn^2$ for all $n \geq n_0$. We claim that $n \geq n_0$ implies the number of rationals $x \in I$ with denominator at most n is at least cn^2 .

To do this, consider the injection $F_{\lfloor n/b \rfloor} \rightarrow I$ given by $x \mapsto (a+x)/b$. Note that the rationals in the image of this map have denominator at most n . Hence the total number of rationals in I with denominator at most n is at least the order of $F_{\lfloor n/b \rfloor}$. For all $n \geq n_0$, this order is at least $c_1(\lfloor n/b \rfloor)^2$, which in turn is greater than cn^2 . \square

This completes the proof of the Main Theorem in the case where M is a sphere.

3. Proof of the Main Theorem for the torus

Assume now that M is diffeomorphic to the torus and has infinite isometry group. In this case, there exists an isometric covering map from

$$\sigma : (\mathbb{R} \times \mathbb{R}, ds^2 + h(s)^2 d\theta^2) \rightarrow M,$$

where $h : \mathbb{R} \rightarrow \mathbb{R}$ is some smooth, positive, and periodic function on \mathbb{R} , as in Figure 4. To fix notation, we perform a global scaling so that the period is π .

As with the case where M is diffeomorphic to \mathbb{S}^2 , we obtain the same geodesic equations and Clairaut relation. However, Lemma 2.1 does not hold since it is possible for geodesics to have the property $|s(t)| \rightarrow \infty$ as $t \rightarrow \infty$. Indeed, this is the case for meridians. As a substitute, we make the following easy observation.

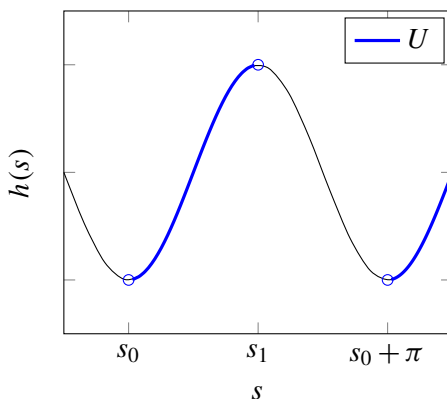


Figure 4. Example of function $h(s)$ corresponding to a torus of revolution. Here, s is the arclength coordinate. The s -values congruent to s_0 or s_1 modulo π correspond to parallel geodesics. The blue region labeled U is, by analogy with the sphere case, the set of oscillating s -values a such that $h'(a) > 0$.

Lemma 3.1. *The π -periodic function $h : \mathbb{R} \rightarrow \mathbb{R}$ has at least one of the two following properties:*

- (1) *(nonisolated case) There exist infinitely many critical points in $(0, \pi)$.*
- (2) *(asymptotic case) There exists an isolated local minimum at some $s_0 \in \mathbb{R}$.*

In the first case of the lemma, it follows that $N(\ell) = \infty$ for all $\ell \geq 2\pi \max(h)$. In the second case, it follows as in the case where M is a sphere that the rotation function $R(a)$ is unbounded. One can imagine why this happens if $h(s)$ is as in Figure 4, since $R(a) \rightarrow \infty$ as $a \rightarrow s_0$ from the right. Given that $R(a)$ is unbounded, it follows that $R(a)$ is not locally constant and hence that $N(\ell) \geq c\ell^2$ asymptotically in ℓ for some constant $c > 0$. This concludes the proof in this case, and it concludes the proof of both theorems in the Introduction.

Acknowledgements

This project began as part of the Summer Undergraduate Research Fellowship program in the College of Creative Studies at UCSB. Rainone is grateful for the support provided by this program. Kennard was partially supported by NSF grants DMS-1045292 and DMS-1404670. Both authors would like to thank Wolfgang Ziller for helpful comments in the preparation of this article.

References

- [Bangert 1993] V. Bangert, “On the existence of closed geodesics on two-spheres”, *Internat. J. Math.* **4**:1 (1993), 1–10. MR Zbl
- [Berger 2010] M. Berger, *Geometry revealed: a Jacob’s ladder to modern higher geometry*, Springer, Heidelberg, 2010. MR Zbl
- [Borzellino et al. 2007] J. E. Borzellino, C. R. Jordan-Squire, G. C. Petrics, and D. M. Sullivan, “Closed geodesics on orbifolds of revolution”, *Houston J. Math.* **33**:4 (2007), 1011–1025. MR Zbl
- [Bröcker and tom Dieck 1985] T. Bröcker and T. tom Dieck, *Representations of compact Lie groups*, Graduate Texts in Mathematics **98**, Springer, New York, 1985. MR Zbl
- [Franks 1992] J. Franks, “Geodesics on S^2 and periodic points of annulus homeomorphisms”, *Invent. Math.* **108**:2 (1992), 403–418. MR Zbl
- [Hingston 1993a] N. Hingston, “Curve shortening, equivariant Morse theory, and closed geodesics on the 2-sphere”, pp. 423–429 in *Differential geometry: Riemannian geometry* (Los Angeles, CA, 1990), edited by R. Greene, Proc. Sympos. Pure Math. **54**, Amer. Math. Soc., Providence, RI, 1993. MR Zbl
- [Hingston 1993b] N. Hingston, “On the growth of the number of closed geodesics on the two-sphere”, *Internat. Math. Res. Notices* **9** (1993), 253–262. MR Zbl
- [Katok 1982] A. Katok, “Entropy and closed geodesics”, *Ergodic Theory Dynam. Systems* **2**:3–4 (1982), 339–365. MR Zbl
- [Katok 1988] A. Katok, “Four applications of conformal equivalence to geometry and dynamics”, *Ergodic Theory Dynam. Systems* **8***:Charles Conley Memorial Issue (1988), 139–152. MR

- [Kobayashi 1972] S. Kobayashi, *Transformation groups in differential geometry*, Ergebnisse der Mathematik und ihrer Grenzgebiete **70**, Springer, New York, 1972. MR Zbl
- [Margulis 1969] G. A. Margulis, “Certain applications of ergodic theory to the investigation of manifolds of negative curvature”, *Funkcional. Anal. i Priložen.* **3:4** (1969), 89–90. In Russian; translated in *Funct. Anal. Appl* **3:4** (1969), 335–336. MR Zbl
- [Oancea 2015] A. Oancea, “Morse theory, closed geodesics, and the homology of free loop spaces”, pp. 67–109 in *Free loop spaces in geometry and topology*, edited by J. Latschev and A. Oancea, IRMA Lect. Math. Theor. Phys. **24**, Eur. Math. Soc., Zürich, 2015. MR
- [Petersen 2016] P. Petersen, *Riemannian geometry*, 3rd ed., Graduate Texts in Mathematics **171**, Springer, Cham, 2016. MR Zbl
- [Walfisz 1963] A. Walfisz, *Weylsche Exponentialsummen in der neueren Zahlentheorie*, Mathematische Forschungsberichte **15**, VEB Deutscher Verlag der Wissenschaften, Berlin, 1963. MR Zbl

Received: 2015-08-30 Revised: 2016-03-07 Accepted: 2016-03-25

kennard@math.ou.edu

*Department of Mathematics, University of Oklahoma,
Norman, OK 73019, United States*

jordan.rainone@stonybrook.edu

*Department of Mathematics, Stony Brook University,
100 Nicolls Road, Stony Brook, NY 11794, United States*

A necessary and sufficient condition for coincidence with the weak topology

Joseph Clanin and Kristopher Lee

(Communicated by Joel Foisy)

For a topological space X , it is a natural undertaking to compare its topology with the weak topology generated by a family of real-valued continuous functions on X . We present a necessary and sufficient condition for the coincidence of these topologies for an arbitrary family $\mathcal{A} \subset C(X)$. As a corollary, we give a new proof of the fact that families of functions which separate points on a compact space induce topologies that coincide with the original topology.

1. Introduction

Given a topological space (X, τ) , let $C(X)$ denote the collection of all continuous functions from X to \mathbb{R} , where \mathbb{R} is equipped with its usual topology. The weak topology induced by a family $\mathcal{A} \subset C(X)$, which we denote by $\tau_{\mathcal{A}}$, is the topology on X such that the collection of sets of the form

$$V(f, y, \epsilon) = \{x \in X : |f(x) - f(y)| < \epsilon\},$$

where $y \in X$, $f \in \mathcal{A}$, and $\epsilon > 0$, is a subbase. It is also characterized as the coarsest topology making all the functions in \mathcal{A} continuous, and thus $\tau_{\mathcal{A}} \subset \tau$. This naturally leads one to ask when equality holds.

Gillman and Jerison [1976, Theorem 3.7] demonstrated that if $\tau = \tau_{\mathcal{A}}$, then the space X is completely regular; however, the converse does not hold in general. For example, if we take (X, τ) to be the real line with the discrete topology and the family \mathcal{A} to consist of only the identity function, then $\tau_{\mathcal{A}}$ is the usual topology on \mathbb{R} and so $\tau_{\mathcal{A}} \neq \tau$.

Conditions for the coincidence of τ and $\tau_{\mathcal{A}}$ are also given. A family $\mathcal{A} \subset C(X)$ is said to be *completely regular* if given a closed set $F \subset X$ and a point $x_0 \in X \setminus F$, there exists an $f \in \mathcal{A}$ with $f(x_0) \notin \text{cl } f[F]$. It is known (see [Gillman and Jerison 1976, Problem 3H]) that if \mathcal{A} is completely regular, then $\tau = \tau_{\mathcal{A}}$. The converse also fails to hold, as we will demonstrate with Example 1.

MSC2010: 46E25, 54A10.

Keywords: weak topology, continuous functions.

At present, a condition that is both necessary *and* sufficient appears to be absent from the literature. To remedy this lapse, we propose the following improvement to the definition of completely regular family:

Definition. A family $\mathcal{A} \subset C(X)$ is said to be finitely completely regular if given a closed set $F \subset X$ and a point $x_0 \in X \setminus F$, there exist $f_1, \dots, f_n \in \mathcal{A}$ such that $0 \notin \text{cl } g[F]$, where the map $g : X \rightarrow \mathbb{R}$ is defined by

$$g(x) = \max_{1 \leq k \leq n} |f_k(x) - f_k(x_0)|.$$

We will show that the condition of finite complete regularity is both necessary and sufficient for $\tau_{\mathcal{A}}$ and τ to coincide, discuss the implications of our result for families \mathcal{A} on compact spaces, and present examples.

2. Main theorem

Theorem. Let (X, τ) be a topological space and let $\mathcal{A} \subset C(X)$ be a family of real-valued continuous functions on X . The weak topology generated by \mathcal{A} coincides with τ if and only if \mathcal{A} is a finitely completely regular family.

Proof. Suppose $\tau = \tau_{\mathcal{A}}$, let F be closed, and let $x_0 \notin F$. As the collection $V(f, y, \epsilon)$ forms a subbase for $\tau_{\mathcal{A}}$, there exist $f_1, \dots, f_n \in \mathcal{A}$ and an $\epsilon > 0$ such that

$$x_0 \in \bigcap_{k=1}^n V(f_k, x_0, \epsilon) \subset X \setminus F,$$

and taking the complement yields

$$F \subseteq \bigcup_{k=1}^n X \setminus V(f_k, x_0, \epsilon).$$

Each set $X \setminus V(f_k, x_0, \epsilon)$ consists of all points $x \in X$ such that $|f_k(x) - f_k(x_0)| \geq \epsilon$, and so if $g : X \rightarrow \mathbb{R}$ is defined by $g(x) = \max\{|f_k(x) - f_k(x_0)| : 1 \leq k \leq n\}$, then $0 \notin \text{cl } g(X \setminus V(f_k, x_0, \epsilon))$ for each k . Therefore, as

$$\text{cl } g(F) \subseteq \bigcup_{k=1}^n \text{cl } g(X \setminus V(f_k, x_0, \epsilon)),$$

we have $0 \notin \text{cl } g(F)$ and thus the family \mathcal{A} is finitely completely regular.

Now, let \mathcal{A} be a finitely completely regular family. Given $U \in \tau$ and $x_0 \in U$, there exist $f_1, \dots, f_n \in \mathcal{A}$ such that $0 \notin \text{cl } g(X \setminus U)$, where $g(x) = \max |f_k(x) - f_k(x_0)|$. Consequently, there exists an $\epsilon > 0$ such that $g(x) \geq \epsilon$ for all $x \in X \setminus U$, and we have

$$X \setminus U \subseteq \bigcup_{i=1}^n \{x \in X : |f_i(x) - f_i(x_0)| \geq \epsilon\},$$

which we complement to obtain

$$x_0 \in \bigcap_{i=1}^n \{x \in X : |f_i(x) - f_i(x_0)| < \epsilon\} \subseteq U.$$

Therefore $\tau \subset \tau_A$, and so $\tau = \tau_A$. □

A family $\mathcal{A} \subset C(X)$ is said to *separate points* if for all distinct $x, y \in X$ there exists a function $f \in \mathcal{A}$ such that $f(x) \neq f(y)$. It is well known that if a family separates points on a compact space, then $\tau_A = \tau$ (see [Kaniuth 2009, Proposition 2.2.14], among others). The main theorem yields a new proof of this fact:

Corollary. *Let (X, τ) be a compact space. If $\mathcal{A} \subset C(X)$ is a family of functions that separates points then $\tau = \tau_A$.*

Proof. We proceed by contraposition. Indeed, suppose $\tau \neq \tau_A$. Then \mathcal{A} fails to be finitely completely regular. Consequently, there exists a closed F and a point $x_0 \in X \setminus F$ such that $0 \in \text{cl } g[F]$, where $g(x) = \max |f_k(x) - f_k(x_0)|$ for any finite collection $f_1, \dots, f_n \in \mathcal{A}$. Since X is compact, g is a closed mapping and this implies that $\text{cl } g[F] = g[F]$, which yields $0 \in g[F]$ and so there exists an $x \in F$ with $f_k(x) = f_k(x_0)$ for each $1 \leq k \leq n$.

Define the closed sets

$$F_f = \{x \in F : f(x) = f(x_0)\} \quad \text{and} \quad K = \bigcap_{f \in \mathcal{A}} F_f.$$

As any finite collection of functions $f_1, \dots, f_n \in \mathcal{A}$ satisfies

$$\bigcap_{k=1}^n F_{f_k} \neq \emptyset,$$

the collection of closed sets $\{F_f : f \in \mathcal{A}\}$ has the finite intersection property and so there exists a $y \in K$. By construction, $f(y) = f(x_0)$ for all $f \in \mathcal{A}$ and since $y \in F$, it must be that $y \neq x_0$. Therefore, \mathcal{A} does not separate points. □

3. Examples

We now give illustrative examples of families of continuous functions; one is finitely completely regular and the other fails to satisfy the definition.

Example 1. Consider the two functions $f, g \in C([0, 1])$ shown in Figure 1. The family $\mathcal{A} = \{f, g\}$ separates points, and thus the topology it induces on $[0, 1]$ is the usual topology. This implies that \mathcal{A} is finitely completely regular; however, it is worth noting that \mathcal{A} fails to be completely regular. Indeed, let $F = [0, \frac{1}{9}] \cup [\frac{5}{9}, \frac{2}{3}]$ and $x_0 = \frac{1}{3}$; then $x_0 \notin F$ but $f(x_0) \in \text{cl } f[F]$ and $g(x_0) \in \text{cl } g[F]$.

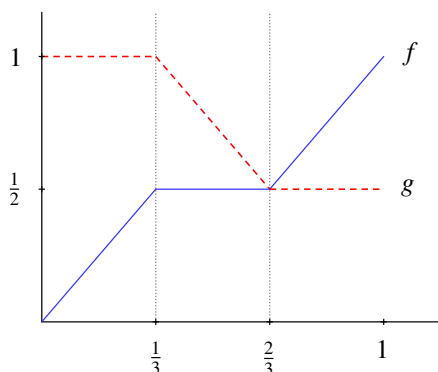


Figure 1. The family $\{f, g\}$ is finitely completely regular, but not completely regular.

It is interesting to note that the subfamily $\{f\}$ of the family in Figure 1 is not finitely completely regular because any interval of the form $(\frac{1}{3} + \epsilon, \frac{2}{3} - \epsilon)$ for $0 < \epsilon < \frac{1}{6}$ is open in the usual topology of the unit interval, but not in the weak topology induced by $\{f\}$. The next example gives a family on $[0, \infty)$ that does not induce a topology that coincides with that of the original space.

Example 2. Let $\mathcal{A} = \{f(x) = \alpha x e^{-x} : \alpha \in \mathbb{R}^+\} \subset C([0, \infty))$, $F = [1, \infty)$, and $x_0 = 0$. For any finite collection $f_1, \dots, f_n \in \mathcal{A}$, where $f_k(x) = \alpha_k x e^{-x}$, we have $0 \in \text{cl } g(F)$, as $g(x) = \max |f_k(x) - f_k(x_0)| = \alpha_j x e^{-x}$ for some $1 \leq j \leq n$. Consequently, \mathcal{A} fails to be finitely completely regular and so $\tau_{\mathcal{A}}$ is strictly coarser than the usual topology on $[0, \infty)$. See Figure 2 for an example.

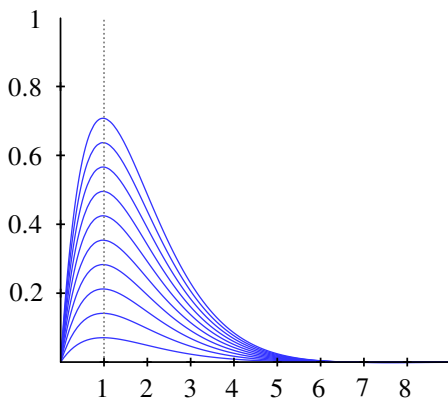


Figure 2. The finite collection $\{f_n(x) = \frac{1}{5}n x e^{-x} : n = 1, \dots, 10\} \subset \mathcal{A}$. Note that $f_n(x) \rightarrow 0$ as $x \rightarrow \infty$ for each $1 \leq n \leq 10$, and this forces $0 \in \text{cl } g[[1, \infty)]$, where $g(x) = \max_{1 \leq k \leq 10} |f_k(x) - f_k(0)|$.

4. Concluding remarks

In this work we have given necessary and sufficient conditions for the coincidence of a topology and a weak topology induced by a family of continuous functions. In particular, this characterization yields a new, more direct proof of the fact that a family that separates points on a compact space will induce the original topology. The definition we introduce additionally reveals that coincidence of the two topologies is possible only when the functions in the family suitably interact with the topology, and our second example illustrates that this can fail even with uncountably many functions.

References

- [Gillman and Jerison 1976] L. Gillman and M. Jerison, *Rings of continuous functions*, Graduate Texts in Mathematics **43**, Springer, New York, NY, 1976. MR Zbl
- [Kaniuth 2009] E. Kaniuth, *A course in commutative Banach algebras*, Graduate Texts in Mathematics **246**, Springer, New York, NY, 2009. MR Zbl

Received: 2015-09-10

Revised: 2015-12-09

Accepted: 2015-12-19

jsc@iastate.edu

*Department of Mathematics, Iowa State University,
Ames, IA 50014, United States*

leekm@iastate.edu

*Department of Mathematics, Iowa State University,
Ames, IA 50014, United States*

Peak sets of classical Coxeter groups

Alexander Diaz-Lopez, Pamela E. Harris,
 Erik Insko and Darleen Perez-Lavin

(Communicated by Stephan Garcia)

We say a permutation $\pi = \pi_1\pi_2 \cdots \pi_n$ in the symmetric group \mathfrak{S}_n has a *peak* at index i if $\pi_{i-1} < \pi_i > \pi_{i+1}$ and we let $P(\pi) = \{i \in \{1, 2, \dots, n\} \mid i \text{ is a peak of } \pi\}$. Given a set S of positive integers, we let $P(S; n)$ denote the subset of \mathfrak{S}_n consisting of all permutations π where $P(\pi) = S$. In 2013, Billey, Burdzy, and Sagan proved $|P(S; n)| = p(n)2^{n-|S|-1}$, where $p(n)$ is a polynomial of degree $\max(S)-1$. In 2014, Castro-Velez et al. considered the Coxeter group of type B_n as the group of signed permutations on n letters and showed that $|P_B(S; n)| = p(n)2^{2n-|S|-1}$, where $p(n)$ is the same polynomial of degree $\max(S)-1$. In this paper we partition the sets $P(S; n) \subset \mathfrak{S}_n$ studied by Billey, Burdzy, and Sagan into subsets of permutations that end with an ascent to a fixed integer k (or a descent to a fixed integer k) and provide polynomial formulas for the cardinalities of these subsets. After embedding the Coxeter groups of Lie types C_n and D_n into \mathfrak{S}_{2n} , we partition these groups into bundles of permutations $\pi_1\pi_2 \cdots \pi_n \mid \pi_{n+1} \cdots \pi_{2n}$ such that $\pi_1\pi_2 \cdots \pi_n$ has the same relative order as some permutation $\sigma_1\sigma_2 \cdots \sigma_n \in \mathfrak{S}_n$. This allows us to count the number of permutations in types C_n and D_n with a given peak set S by reducing the enumeration to calculations in the symmetric group and sums across the rows of Pascal's triangle.

1. Introduction

We say a permutation $\pi = \pi_1\pi_2 \cdots \pi_n$ in the symmetric group \mathfrak{S}_n has a *peak* at index i if $\pi_{i-1} < \pi_i > \pi_{i+1}$. We let $[n] := \{1, 2, \dots, n\}$ and define the peak set of a permutation π to be the set of peaks in π :

$$P(\pi) = \{i \in [n] \mid i \text{ is a peak of } \pi\}.$$

Given a subset $S \subset [n]$, we denote the set of all permutations with peak set S by

$$P(S; n) = \{\pi \in \mathfrak{S}_n \mid P(\pi) = S\}.$$

MSC2010: 05A05, 05A10, 05A15.

Keywords: binomial coefficient, peak, permutation, signed permutation, permutation pattern.

This research was performed while Harris held a National Research Council Research Associateship Award at USMA/ARL.

We say a set $S \subset [n]$ is *n-admissible* (or simply *admissible* when n is understood) provided $P(S; n) \neq \emptyset$.

While the combinatorics of Coxeter groups has fascinated mathematicians for generations [Björner and Brenti 2005], the combinatorics of peaks has only recently caught the eye of the mathematical community. Stembridge [1997] was one of the first to study the combinatorics of peaks; he defined a peak analog of Stanley's theory of poset partitions. Nyman [2003] showed that taking formal sums of permutations according to their peak sets gives a nonunital subalgebra of the group algebra of the symmetric group. This motivated several papers studying peak (and descent) algebras of classical Coxeter groups [Aguiar et al. 2004; 2006b, Bergeron and Hohlweg 2006; Petersen 2007]. Peaks have also been linked to the Schubert calculus of isotropic flag manifolds [Bergeron et al. 2002; Bergeron and Sottile 2002; Billey and Haiman 1995] and the generalized Dehn–Sommerville equations [Aguiar et al. 2006a; Bergeron et al. 2000; Billera et al. 2003].

Billey, Burdzy, and Sagan [Billey et al. 2013, Theorem 1.1] counted the number of elements in the sets $P(S; n)$. For any n -admissible set S , they found these cardinalities satisfy

$$|P(S; n)| = p(n)2^{n-|S|-1}, \quad (1)$$

where $|S|$ denotes the cardinality of the set S , and where the *peak polynomial* $p(n)$ is a polynomial of degree $\max(S)-1$ that takes integral values when evaluated at integers. Their study was motivated by a problem in probability theory which explored the mass distribution on graphs as it relates to random permutations with specific peak sets; this research was presented in [Billey et al. 2015]. Billey, Burdzy, and Sagan also computed closed formulas for the peak polynomials $p(n)$ for various special cases of $P(S; n)$ using the method of finite differences, and Billey, Fahrback, and Talmage [Billey et al. 2016] then studied the coefficients and zeros of peak polynomials.

Shortly after Billey, Burdzy, and Sagan's article appeared on the arXiv, Kasraoui [2012] proved one of their open conjectures and identified the most probable peak set for a random permutation. Then Castro-Velez et al. [2013] generalized the work of Billey, Burdzy, and Sagan to study peak sets of type- B signed permutations. They studied two sets $P_B(S; n)$ and $\hat{P}_B(S; n)$ of signed permutations with peak set S , whose formal definition we introduce in Section 3B. Their main result regarding the set $P_B(S; n)$ [Castro-Velez et al. 2013, Theorem 2.4] used induction to prove

$$|P_B(S; n)| = |P(S; n)|2^n = p(n)2^{2n-|S|-1}. \quad (2)$$

Note that $p(n)$ is the same polynomial as that of (1).

Motivated by extending the above-mentioned results to other classical Coxeter groups, our work begins by partitioning the sets $P(S; n)$ studied by Billey, Burdzy, and Sagan into subsets $P(S; n)^{\nearrow k}$ and $P(S; n)_{\searrow k}$ of permutations ending with an

ascent or a descent to a fixed k , respectively. With these partitions on hand, we show in Theorems 11 and 12 that the cardinalities of these sets are governed by polynomial formulas similar to those discovered by Billey, Burdzy, and Sagan. These results are presented in Section 2.

We then embed the Coxeter groups of types C_n and D_n into \mathfrak{S}_{2n} and call these embedded subgroups $\mathcal{C}_n, \mathcal{D}_n \subset \mathfrak{S}_{2n}$ the *mirrored permutations* of types C_n and D_n , respectively (Section 3). For each $\pi \in \mathfrak{S}_n$, we define the *pattern bundle* of π in types C_n and D_n in Definitions 14 and 17. Each *pattern bundle* consists of permutations $\tau_1 \tau_2 \cdots \tau_n \mid \tau_{n+1} \cdots \tau_{2n}$ such that $\tau_1 \tau_2 \cdots \tau_n$ flattens to $\pi_1 \pi_2 \cdots \pi_n$, meaning $\tau_1 \tau_2 \cdots \tau_n$ has the same relative order as $\pi_1 \pi_2 \cdots \pi_n$. These pattern bundles have the following properties: (1) they partition the groups \mathcal{C}_n and \mathcal{D}_n ; (2) they are indexed by the elements of \mathfrak{S}_n , and; (3) they have size 2^n in \mathcal{C}_n and 2^{n-1} in \mathcal{D}_n . This process allows us to give concise proofs of the following two identities (Theorem 24(I) and (II), respectively):

$$|P_C(S; n)| = p(n)2^{2n-|S|-1} \quad \text{and} \quad |P_D(S; n)| = p(n)2^{2n-|S|-2}.$$

We note that the polynomial appearing above is the same as that of (1). Moreover, the proof of Theorem 24(I) is much shorter than the one given by [Castro-Velez et al. 2013, Theorem 2.4], and Theorem 24(II) has not appeared before in the literature.

Finally in Section 4 we prove our main result, Theorem 26. We use the formulas for $|P(S; n)^{\nearrow k}|$ and $|P(S; n)_{\searrow k}|$ from Section 2 and sums of binomial coefficients to enumerate the set of permutations with peak set $S \subset [n]$ in \mathcal{C}_n and \mathcal{D}_n .

We end this introduction with a remark on the history of this collaboration. The last three authors of this article began their study of peak sets in classical Coxeter groups before Castro-Velez et al. had published their results from type B_n , and focused their study on the Coxeter (Weyl) groups of types C_n and D_n using presentations of these groups described in [Billey and Lakshmibai 2000, pp. 29, 34]. While Perez-Lavin was presenting the preliminary results of this paper at the USTARS 2014 conference held at UC Berkeley, we met Alexander Diaz-Lopez, who told us of his recently completed work with Castro-Velez et al. [2013]. Knowing that the Coxeter groups of types B and C are isomorphic, we were immediately intrigued to see what connections could be found between the two works. We were delighted to find that we used vastly different techniques to count the elements of $P_B(S; n)$ and $P_C(S; n)$, and discovered an isomorphism between the two groups which preserves peak sets (up to a reordering of the peaks). We highlight these connections and compare and contrast the two works in Section 3B.

2. Partitioning the set $P(S; n)$

To make our approach precise, we begin by setting notation and giving some definitions.

Definition 1. For a given peak set $S \subset [n - 1]$, we define

$$P(S; n)^{\nearrow k} := \{ \pi \in P(S; n) \mid \pi_{n-1} < \pi_n \text{ and } \pi_n = k \}, \quad \overline{P(S; n)} := \bigsqcup_{k=1}^n P(S; n)^{\nearrow k},$$

$$P(S; n)_{\searrow k} := \{ \pi \in P(S; n) \mid \pi_{n-1} > \pi_n \text{ and } \pi_n = k \}, \quad \underline{P(S; n)} := \bigsqcup_{k=1}^n P(S; n)_{\searrow k}.$$

We remark that $P(S; n)^{\nearrow 1} = \emptyset$ because a permutation cannot end with an ascent to 1. Similarly $P(S; n)_{\searrow n} = \emptyset$ since a permutation cannot end with a descent to n . Therefore the sets $\overline{P(S; n)}$ and $\underline{P(S; n)}$ are the disjoint unions of sets

$$\overline{P(S; n)} = \bigsqcup_{k=2}^n P(S; n)^{\nearrow k} \quad \text{and} \quad \underline{P(S; n)} = \bigsqcup_{k=1}^{n-1} P(S; n)_{\searrow k}.$$

Since every $\pi \in P(S; n)$ either ends with an ascent or a descent, we see

$$P(S; n) = \overline{P(S; n)} \sqcup \underline{P(S; n)}.$$

Our next lemma counts the permutations without peaks that end with an ascent to k .

Lemma 2. *If $2 \leq k \leq n$, then $|P(\emptyset; n)^{\nearrow k}| = 2^{k-2}$.*

Proof. Let $2 \leq k \leq n$ and suppose $\pi = \pi_1 \pi_2 \cdots \pi_n \in P(\emptyset; n)^{\nearrow k}$. Hence $P(\pi) = \emptyset$ and $\pi_{n-1} < \pi_n = k$. Let us further assume that $\pi = \tau_A 1 \tau_B k$, where τ_A and τ_B are the portions of π to the left and right of 1, respectively. Since $P(\pi) = \emptyset$, we know τ_A must decrease, while τ_B must increase. However, the values of τ_B must come from the set $\{2, 3, \dots, k - 1\}$ because $\pi_{n-1} < \pi_n = k$, and there is one $\pi \in P(\emptyset; n)^{\nearrow k}$ for each subset of $\{2, 3, \dots, k - 1\}$ as such a π is completely determined by which elements from that set appear in τ_B . Hence we see $|P(\emptyset; n)^{\nearrow k}| = 2^{k-2}$. \square

We will next prove a recursive formula for the number of permutations with specified peak set S that end in an ascent to a fixed integer k .

Lemma 3. *Let $S \subset [n - 1]$ be a nonempty admissible set. Let $m = \max S$ and fix an integer k , where $1 \leq k \leq n$. If $S_1 = S \setminus \{m\}$ and $S_2 = S_1 \cup \{m - 1\}$, then*

$$|P(S; n)^{\nearrow k}| = \sum_{i=0}^{k-2} \binom{k-1}{i} \binom{n-k}{m-i-1} |P(S_1; m-1)| 2^{k-i-2} - |P(S_1; n)^{\nearrow k}| - |P(S_2; n)^{\nearrow k}|.$$

Proof. Observe that if $k = 1$, then the result holds trivially as all terms in the statement are identically zero. Let $2 \leq k \leq n$ and let $\Pi^{\nearrow k}$ denote the set of permutations ending with an ascent to k that have peak set S_1 in the first $m - 1$ spots and no peaks in the last $m - n + 1$, i.e.,

$$\Pi^{\nearrow k} = \{ \pi \in \mathfrak{S}_n \mid P(\pi_1 \pi_2 \cdots \pi_{m-1}) = S_1, P(\pi_m \cdots \pi_n) = \emptyset, \text{ and } \pi_{n-1} < \pi_n = k \}.$$

We compute the cardinality of the set $\Pi^{\nearrow k}$ by counting the number of ways to construct a permutation in $\Pi^{\nearrow k}$.

First we select a subset $P_1 = \{\pi_1, \pi_2, \dots, \pi_{m-1}\} \subset [n] \setminus \{k\}$ (as we fix π_n to be k). When selecting P_1 , we can choose i numbers from $\{1, 2, \dots, k-1\}$ to include in P_1 for each $0 \leq i \leq k-1$ and then choose the remaining $m-i-1$ numbers from the set $\{k+1, k+2, \dots, n\}$ to fill the remainder of P_1 . Thus there are $\binom{k-1}{i} \cdot \binom{n-k}{m-i-1}$ ways to select the elements of P_1 . By definition, there are $|P(S_1, m-1)|$ ways to arrange the $m-1$ elements of P_1 into a permutation $\pi_1\pi_2 \cdots \pi_{m-1}$ satisfying $P(\pi_1\pi_2 \cdots \pi_{m-1}) = S_1$.

Let $P_2 = \{\pi_m, \pi_{m+1}, \dots, \pi_n\} = [n] \setminus P_1$, where $\pi_n = k$. There are $n - (m-1) = n - m + 1$ numbers in P_2 , and there are precisely $k - i - 1$ elements from the set $\{1, 2, \dots, k-1\}$ that were not chosen to be part of P_1 . That means k is the $(k-i)$ -th largest integer in the set P_2 . By flattening the numbers in P_2 , we can see there are $|P(\emptyset; n - m + 1)^{\nearrow k-i}|$ ways to arrange the elements of P_2 to create a subpermutation $\pi_m\pi_{m+1} \cdots \pi_n$ that satisfies

$$P(\pi_m \cdots \pi_n) = \emptyset \quad \text{and} \quad \pi_{n-1} < \pi_n = k.$$

By Lemma 2 we know that $|P(\emptyset; n - m + 1)^{\nearrow k-i}| = 2^{k-i-2}$ when $k - i \geq 2$ and it is 0 otherwise. Of course $k - i \geq 2$ when $i \leq k - 2$. Putting this all together, we see that the number of ways to create a permutation in $\Pi^{\nearrow k}$ is

$$|\Pi^{\nearrow k}| = \sum_{i=0}^{k-2} \binom{k-1}{i} \binom{n-k}{m-i-1} |P(S_1; m-1)| 2^{k-i-2}. \tag{3}$$

Next we consider a different way to count the elements of $\Pi^{\nearrow k}$. Note that we have not specified whether $\pi_{m-1} > \pi_m$ or $\pi_{m-1} < \pi_m$. So, in particular, based on the definition of $\Pi^{\nearrow k}$ and its restrictions on $P(\pi_1\pi_2 \cdots \pi_{m-1})$ and $P(\pi_m\pi_{m+1} \cdots \pi_n)$, all of the following are possible:

$$P(\pi) = S, \quad P(\pi) = S_1, \quad \text{or} \quad P(\pi) = S_2 \quad \text{for } \pi \in \Pi^{\nearrow k}.$$

Hence

$$\Pi^{\nearrow k} = P(S; n)^{\nearrow k} \sqcup P(S_1; n)^{\nearrow k} \sqcup P(S_2; n)^{\nearrow k}.$$

Thus

$$|\Pi^{\nearrow k}| = |P(S; n)^{\nearrow k}| + |P(S_1; n)^{\nearrow k}| + |P(S_2; n)^{\nearrow k}|. \tag{4}$$

The result follows from setting (3) and (4) equal to each other and solving for the quantity $|P(S; n)^{\nearrow k}|$. □

The following lemma will be used in the proofs of Lemmas 5 and 9.

Lemma 4. *If $n \geq 2$ then*

- $|P(\emptyset; n)| = 1$, and
- $|P(\emptyset; n)| = 2^{n-1} - 1$.

Proof. The only permutation $\pi \in P(\emptyset; n)$ that ends in a descent is $n = \pi_1 > \pi_2 > \dots > \pi_n = 1$; therefore $|P(\emptyset; n)| = 1$. On the other hand, it is easy to see that $P(\emptyset; n) = 2^{n-1}$, as in [Billiey et al. 2013, Proposition 2.1]. Since $P(\emptyset; n) = \overline{P(\emptyset; n)} \sqcup \underline{P(\emptyset; n)}$, we compute

$$|\overline{P(\emptyset; n)}| = |P(\emptyset; n)| - |\underline{P(\emptyset; n)}| = 2^{n-1} - 1. \quad \square$$

The following result allows us to recursively enumerate the set of permutations with specified peak set S that end with an ascent.

Lemma 5. *Let $S \subset [n - 1]$ be a nonempty n -admissible set, and let $m = \max S$. If we let $S_1 = S \setminus \{m\}$ and $S_2 = S_1 \cup \{m - 1\}$, then*

$$|\overline{P(S; n)}| = \binom{n}{m-1} (2^{n-m} - 1) |P(S_1; m - 1)| - |\overline{P(S_1; n)}| - |\overline{P(S_2; n)}|.$$

Proof. Let $S \subset [n - 1]$ be an admissible set with $m = \max S$. Define the sets $S_1 = S \setminus \{m\}$, $S_2 = S_1 \cup \{m - 1\}$ and

$$\Pi^{\nearrow} = \{ \pi \in \mathfrak{S}_n \mid P(\pi_1 \pi_2 \dots \pi_{m-1}) = S_1, P(\pi_m \dots \pi_n) = \emptyset, \text{ and } \pi_{n-1} < \pi_n \}.$$

Next we compute the cardinality of Π^{\nearrow} . We observe that there are $\binom{n}{m-1}$ choices for the values of π_1, \dots, π_{m-1} , and by definition, there are $|P(S_1; m - 1)|$ ways to arrange the values of π_1, \dots, π_{m-1} so that $P(\pi_1 \pi_2 \dots \pi_{m-1}) = S_1$. Once we have chosen the values of $\pi_1, \pi_2, \dots, \pi_{m-1}$, the values of

$$\pi_m, \pi_{m+1}, \pi_{m+2}, \dots, \pi_n$$

are determined. We note that there are $|\overline{P(\emptyset; n - m + 1)}|$ ways to arrange the values of π_m, \dots, π_n , so that $P(\pi_m \dots \pi_n) = \emptyset$ and $\pi_{n-1} < \pi_n$.

Yet Lemma 4 proved that $|\overline{P(\emptyset; n - m + 1)}| = 2^{n-m} - 1$. Hence we see that

$$|\Pi^{\nearrow}| = \binom{n}{m-1} (2^{n-m} - 1) |P(S_1; m - 1)|. \quad (5)$$

On the other hand $\Pi^{\nearrow} = \overline{P(S; n)} \sqcup \overline{P(S_1; n)} \sqcup \overline{P(S_2; n)}$ by the defining conditions of Π^{\nearrow} . Hence

$$|\Pi^{\nearrow}| = |\overline{P(S; n)}| + |\overline{P(S_1; n)}| + |\overline{P(S_2; n)}|. \quad (6)$$

When we set the right-hand sides of (5) and (6) equal to each other and solve for $|\overline{P(S; n)}|$, we see that

$$|\overline{P(S; n)}| = \binom{n}{m-1} (2^{n-m} - 1) |P(S_1; m - 1)| - |\overline{P(S_1; n)}| - |\overline{P(S_2; n)}|. \quad \square$$

The following examples illustrate the recursion used to prove Lemmas 3 and 5.

Example 6. We make use of Lemma 3 to compute $|P(\{3\}; 5)^{\nearrow 3}|$. Let S be the set $S = \{3\} \subset [5]$. Note that $m = \max S = 3$. Then we compute

$$|P(\{3\}; 5)^{\nearrow 3}| = \left[\binom{2}{0} \binom{2}{2} 2^1 + \binom{2}{1} \binom{2}{1} 2^0 \right] |P(\emptyset; 2)| - |P(\emptyset; 5)^{\nearrow 3}| - |P(\{2\}; 5)^{\nearrow 3}|. \quad (7)$$

Some small computations show that

$$P(\emptyset; 2) = \{12, 21\}, \quad P(\emptyset; 5)^{\nearrow 3} = \{54213, 54123\}, \\ P(\{2\}; 5)^{\nearrow 3} = \{45213, 25413, 45123, 15423\}.$$

Accordingly, we can see that (7) gives

$$|P(\{3\}; 5)^{\nearrow 3}| = (2 + 4)(2) - 2 - 4 = 6.$$

Example 7. In this example we make use of Lemma 5 to compute $|\overline{P(\{3\}; 5)}|$. If we let $S = \{3\} \subset [5]$ then $m = \max S = 3$. We then have

$$|\overline{P(\{3\}; 5)}| = \binom{5}{2} (2^{5-3} - 1) |P(\emptyset; 2)| - |\overline{P(\emptyset; 5)}| - |\overline{P(\{2\}; 5)}|. \quad (8)$$

Some small computations show

$$P(\emptyset; 2) = \{12, 21\}, \\ \overline{P(\emptyset; 5)} = \left\{ \begin{array}{l} 54321, 54213, 54123, 53214, 53124, 52134, 51234, 43215, \\ 43125, 42135, 32145, 41235, 31245, 21345, 12345 \end{array} \right\}.$$

Direct computations yield

$$\overline{P(\{2\}; 5)} \\ = \left\{ \begin{array}{l} 45312, 35412, 45213, 25413, 45123, 15423, 35214, 25314, 35124, 25134, \\ 15324, 15234, 34215, 24315, 34125, 24135, 23145, 14325, 14235, 13245 \end{array} \right\}.$$

Equation (8) gives

$$|\overline{P(\{3\}; 5)}| = \binom{5}{2} (2^{5-3} - 1)(2) - 15 - 20 = 25.$$

Next we consider permutations that end in a descent to a specific value k .

Lemma 8. Let $S \subset [n - 1]$ be a nonempty admissible set, let $m = \max S$, and fix an integer k , where $1 \leq k \leq n$. If $S_1 = S \setminus \{m\}$ and $S_2 = S_1 \cup \{m - 1\}$, then

$$|P(S; n)_{\searrow k}| = \binom{n-k}{n-m} |P(S_1; m - 1)| - |P(S_1; n)_{\searrow k}| - |P(S_2; n)_{\searrow k}|.$$

The proof of Lemma 8 follows similarly to that of Lemma 3; hence we omit the argument, but point the interested reader to the preprint version of this paper for a detailed proof [Diaz-Lopez et al. 2015].

The following result allows us to recursively enumerate the set of permutations with specified peak set S that end with a descent.

Lemma 9. *Let $S \subset [n - 1]$ be a nonempty admissible set, and let $m = \max S$. If $S_1 = S \setminus \{m\}$ and $S_2 = S_1 \cup \{m - 1\}$, then*

$$|\underline{P}(S; n)| = \binom{n}{m-1} |P(S_1; m - 1)| - |\underline{P}(S_1; n)| - |\underline{P}(S_2; n)|.$$

Proof. By Definition 1,

$$|\underline{P}(S; n)| = \sum_{k=1}^{n-1} |P(S; n)_{\searrow k}|.$$

Using this equation and Lemma 8 we get

$$\begin{aligned} |\underline{P}(S; n)| &= \sum_{k=1}^{n-1} \left[\binom{n-k}{n-m} |P(S_1; m - 1)| - |P(S_1; n)_{\searrow k}| - |P(S_2; n)_{\searrow k}| \right] \\ &= \binom{n}{m-1} |P(S_1; m - 1)| - |\underline{P}(S_1; n)| - |\underline{P}(S_2; n)|, \end{aligned}$$

where the last equality comes from the identity $\sum_{k=0}^n \binom{k}{c} = \binom{n+1}{c+1}$. □

As before, we provide an example that illustrates the use of the previous results.

Example 10. Consider the set $S = \{3\} \subset [5]$; hence $m = \max S = 3$. We want to compute $|P(\{3\}; 5)_{\searrow 2}|$. By Lemma 8 we have

$$|P(\{3\}; 5)_{\searrow 2}| = \binom{3}{2} |P(\emptyset; 2)| - |P(\emptyset; 5)_{\searrow 2}| - |P(\{2\}; 5)_{\searrow 2}|.$$

Some simple computations show that

$$P(\emptyset; 2) = \{12, 21\}, \quad P(\emptyset; 5)_{\searrow 2} = \emptyset, \quad \text{and} \quad P(\{2\}; 5)_{\searrow 2} = \{15432\}.$$

Therefore

$$|P(\{3\}; 5)_{\searrow 2}| = 3(2) - 0 - 1 = 5.$$

In fact, $P(\{3\}; 5)_{\searrow 2} = \{51432, 41532, 31542, 14532, 13542\}$.

We want to compute $|\underline{P}(\{3\}; 5)|$. By Lemma 9 we have

$$|\underline{P}(\{3\}; 5)| = \binom{5}{2} |P(\emptyset; 2)| - |\underline{P}(\emptyset; 5)| - |\underline{P}(\{2\}; 5)|.$$

Again we can compute that

$$\begin{aligned} P(\emptyset; 2) &= \{12, 21\}, & \underline{P}(\emptyset; 5) &= \{54321\}, \\ \underline{P}(\{2\}; 5) &= \{45321, 35421, 25431, 15432\}. \end{aligned}$$

Thus

$$|\underline{P}(\{3\}; 5)| = 10(2) - 1 - 4 = 15.$$

In fact,

$$\underline{P}(\{3\}; 5) = \left\{ \begin{array}{l} 53421, 43521, 34521, 52431, 42531, 32541, 24531, \\ 23541, 51432, 41532, 31542, 21543, 14532, 13542, 12543 \end{array} \right\}.$$

The following two theorems allow us to easily calculate closed formulas for $|\underline{P}(S; n)|$ and $|\underline{P}(S; n)^{\nearrow k}|$ using the method of finite differences [Stanley 2012, Proposition 1.9.2]. We start by applying Lemma 9 in an induction argument to show $|\underline{P}(S; n)|$ is given by a polynomial $p_\delta(n)$.

Theorem 11. *Let $S \subset [n - 1]$ be an admissible set. If $S = \emptyset$, take $m = 1$; otherwise let $m = \max S$. Then the cardinality of the set $\underline{P}(S; n)$ is given by*

$$|\underline{P}(S; n)| = p_\delta(n),$$

where $p_\delta(n)$ is a polynomial in the variable n of degree $m - 1$ that returns integer values for all integers n .

Proof. We induct on the sum $i = \sum_{i \in S} i$. When $i = 0$, the set S is empty. By Lemma 4 we get $|\underline{P}(\emptyset; n)| = 1$, and so $p_\delta(n) = 1$ is a polynomial of degree 0.

Let $S \subset [n - 1]$ be nonempty, with $m = \max S$ and $\sum_{i \in S} i = i$. Let $S_1 = S \setminus \{m\}$ and $S_2 = S_1 \cup \{m - 1\}$, and note, in particular, that the sums $\sum_{i \in S_1} i$ and $\sum_{i \in S_2} i$ are both strictly less than i . By induction, we know $|\underline{P}(S_1; n)| = p_{\delta_1}(n)$ and $|\underline{P}(S_2; n)| = p_{\delta_2}(n)$, where p_{δ_1} and p_{δ_2} are polynomials of degrees less than $m - 1$ that have integral values when evaluated at integers.

By (1) we have $|\underline{P}(S_1; m - 1)| = p(m - 1)2^{(m-1)-|S_1|-1}$ and this expression returns an integer value when evaluated at any integer $m - 1$ [Billey et al. 2013, Theorem 2.2]. Since the expression $p(m - 1)2^{(m-1)-|S_1|-1}$ is an integer-valued constant with respect to n , we see that $p(m - 1)2^{(m-1)-|S_1|-1} \binom{n}{m-1}$ is a polynomial in the variable n of degree $m - 1$. These facts, together with Lemma 9, imply

$$\begin{aligned} |\underline{P}(S; n)| &= \binom{n}{m-1} |\underline{P}(S_1; m - 1)| - |\underline{P}(S_1; n)| - |\underline{P}(S_2; n)| \\ &= \binom{n}{m-1} p(m - 1)2^{(m-1)-|S_1|-1} - p_{\delta_1} - p_{\delta_2} \\ &= p_\delta, \end{aligned}$$

where p_δ is a polynomial in the variable n of degree $m - 1$ that has integer values when evaluated at integers. □

Using Lemma 3, we show $|\underline{P}(S; n)^{\nearrow k}|$ is given by a polynomial.

Theorem 12. *Let $S \subset [n - 1]$ be an admissible set. If $S = \emptyset$ take $m = 1$; otherwise let $m = \max S$. Fix an integer k satisfying $2 \leq k \leq n$; then the cardinality of the set $\underline{P}(S; n)^{\nearrow k}$ is given by*

$$|\underline{P}(S; n)^{\nearrow k}| = p_{\alpha(k)}(n),$$

where $p_{\alpha(k)}(n)$ is a polynomial of degree $m - 1$ that returns integer values for all integers n .

Proof. We proceed by induction on the sum $i = \sum_{i \in S} i$. When $i = 0$ the set S is empty. By Lemma 2 we get $|P(\emptyset; n)^{\nearrow k}| = 2^{k-2}$, which is a polynomial of degree 0 in the indeterminate n .

Consider a nonempty subset $S \subset [n - 1]$ with $m = \max S$ and $\sum_{i \in S} i = i$. Let $S_1 = S \setminus \{m\}$ and $S_2 = S_1 \cup \{m - 1\}$, and note, in particular, that the sums $\sum_{i \in S_1} i$ and $\sum_{i \in S_2} i$ are both strictly less than i . By induction, we know

$$|P(S_1; n)^{\nearrow k}| = p_{\alpha_1(k)}(n) \quad \text{and} \quad |P(S_2; n)^{\nearrow k}| = p_{\alpha_2(k)}(n),$$

where $p_{\alpha_1(k)}(n)$ and $p_{\alpha_2(k)}(n)$ are each polynomials of degrees less than $m - 1$ that have integer values when evaluated at integers.

By (1) we know $|P(S_1; m - 1)| = p(m - 1)2^{(m-1)-|S_1|-1}$ is an integer-valued function when evaluated at any integer $m - 1$, and it is a constant function with respect to n . Hence the expression $\binom{k-1}{i} |P(S_1; m - 1)| 2^{k-i-2}$ is a polynomial expression in n that has degree $m - 1$ when $i = 0$ and degree less than or equal to $m - 1$ for $1 \leq i \leq k - 2$. These facts, together with Lemma 3, imply

$$\begin{aligned} |P(S; n)^{\nearrow k}| &= \sum_{i=0}^{k-2} \binom{k-1}{i} \binom{n-k}{m-i-1} |P(S_1; m-1)| 2^{k-i-2} \\ &\quad - |P(S_1; n)^{\nearrow k}| - |P(S_2; n)^{\nearrow k}| \\ &= \sum_{i=0}^{k-2} \binom{k-1}{i} \binom{n-k}{m-i-1} |P(S_1; m-1)| 2^{k-i-2} - p_{\alpha_1(k)}(n) - p_{\alpha_2(k)}(n) \\ &= p_{\alpha(k)}(n) \end{aligned}$$

is a polynomial in the variable n of degree $m - 1$ that returns integer values when evaluated at integers. □

Below we show an example of how to find the polynomial $p_{\alpha(k)}(n)$.

Example 13. It is well known that any sequence given by a polynomial of degree d can be completely determined by any consecutive $d + 1$ values by the method of finite differences [Stanley 2012, Proposition 1.9.2]. Theorems 11 and 12 give us a way of finding explicit formulas $p_{\alpha(k)}(n)$ and $p_{\delta}(n)$ for an admissible set S .

For instance if $S = \{2, 4\}$ and $k = 6$, Theorem 12 tells us $p_{\alpha(k)}(n)$ is a polynomial of degree 3. Hence we require four consecutive terms to compute $p_{\alpha(k)}(n)$. One can compute that the first few values of $p_{\alpha(6)}(n) = |P(S; n)^{\nearrow 6}|$ are

$$\begin{aligned} p_{\alpha(6)}(6) &= 16, & p_{\alpha(6)}(7) &= 80, & p_{\alpha(6)}(8) &= 224, \\ p_{\alpha(6)}(9) &= 480, & p_{\alpha(6)}(10) &= 880, & p_{\alpha(6)}(11) &= 1456, \quad \dots \end{aligned}$$

We then take four successive differences until we get a row of zeros in the following array:

$$\begin{array}{cccccc}
 16 & 80 & 224 & 480 & 880 & 1456 & \dots \\
 64 & 144 & 256 & 400 & 576 & \dots & \\
 80 & 112 & 144 & 176 & \dots & & \\
 32 & 32 & 32 & \dots & & & \\
 0 & 0 & \dots & & & &
 \end{array}$$

Since the first value we considered in the first row of the array above is the value of $p_{\alpha(6)}(n)$ at $n = 6$, we can write the polynomial $p_{\alpha(6)}(n)$ in the basis $\binom{n-6}{j}$ as

$$p_{\alpha(6)}(n) = 16\binom{n-6}{6} + 64\binom{n-6}{7} + 80\binom{n-6}{8} + 32\binom{n-6}{9}.$$

The sequence given by $\frac{1}{16}p_{\alpha(6)}(n)$ in this example is sequence A000330 in Sloane’s *On-line encyclopedia of integer sequences* [OEIS 1996] with the index n shifted by 6.

3. Pattern bundles of Coxeter groups of types C and D

In this section, we describe embeddings of the Coxeter groups of types C_n and D_n into the symmetric group \mathfrak{S}_{2n} . We then partition these groups into subsets, which we call *pattern bundles* and denote by $\mathcal{C}_n(\pi)$ and $\mathcal{D}_n(\pi)$, that correspond to permutations π of \mathfrak{S}_n . Each of the *type- C_n pattern bundles* $\mathcal{C}_n(\pi)$ contains 2^n elements, and the *type- D_n pattern bundles* $\mathcal{D}_n(\pi)$ contain 2^{n-1} elements. These sets allow us to give a concise proof of Theorem 24, and they play an instrumental role in our proof of Theorem 26.

3A. Pattern bundle algorithms for \mathcal{C}_n and \mathcal{D}_n . We define the group of type- C_n *mirrored permutations* to be the subgroup $\mathcal{C}_n \subset \mathfrak{S}_{2n}$ consisting of all permutations $\pi_1\pi_2 \cdots \pi_n \mid \pi_{n+1}\pi_{n+2} \cdots \pi_{2n} \in \mathfrak{S}_{2n}$, where $\pi_i = k$ if and only if $\pi_{2n-i+1} = 2n - k + 1$. In other words, if we place a “mirror” between π_n and π_{n+1} , then the numbers i and $2n - i + 1$ must be the same distance from the mirror for each $1 \leq i \leq n$. A simple transposition s_i with $1 \leq i \leq n - 1$ acts on a mirrored permutation $\pi \in \mathcal{C}_n \subset \mathfrak{S}_{2n}$ (on the right) by simultaneously transposing π_i with π_{i+1} and π_{2n-i} with π_{2n-i+1} . The simple transposition s_n acts on a mirrored permutation $\pi \in \mathcal{C}_n \subset \mathfrak{S}_{2n}$ by transposing π_n with π_{n+1} .

Similarly, we define the group of type- D_n *mirrored permutations* as the subgroup $\mathcal{D}_n \subset \mathfrak{S}_{2n}$ consisting of all permutations $\pi_1\pi_2 \cdots \pi_n \mid \pi_{n+1}\pi_{n+2} \cdots \pi_{2n} \in \mathfrak{S}_{2n}$, where $\pi_i = k$ if and only if $\pi_{2n-i+1} = 2n - k + 1$ and the set $\{\pi_1, \pi_2, \dots, \pi_n\}$ always contains an even number of elements from the set $\{n + 1, n + 2, \dots, 2n\}$. A simple transposition s_i with $1 \leq i \leq n - 1$ acts on a mirrored permutation $\pi \in \mathcal{D}_n \subset \mathfrak{S}_{2n}$ (on the right) by simultaneously transposing π_i with π_{i+1} and π_{2n-i} with π_{2n-i+1} . The simple transposition s_n acts on a mirrored permutation $\pi \in \mathcal{D}_n \subset \mathfrak{S}_{2n}$ by transposing $\pi_{n-1}\pi_n$ with $\pi_{n+1}\pi_{n+2}$.

Definition 14. Let $\pi = \pi_1\pi_2 \cdots \pi_n \in \mathfrak{S}_n$. We define the *pattern bundle* of π in type C_n (denoted $\mathcal{C}_n(\pi)$) to be the set of all mirrored permutations

$$\tau = \tau_1\tau_2 \cdots \tau_n \mid \tau_{n+1}\tau_{n+2} \cdots \tau_{2n} \in \mathcal{C}_n$$

such that $\tau_1\tau_2 \cdots \tau_n$ has the same relative order as $\pi_1\pi_2 \cdots \pi_n$.

We could equivalently describe $\mathcal{C}_n(\pi)$ as the set of mirrored permutations which contain the *permutation pattern* π in the first n entries. We will show that these sets partition \mathcal{C}_n into subsets of size 2^n . For every $\pi \in \mathfrak{S}_n$, we will describe how to construct the *pattern bundle* $\mathcal{C}_n(\pi) \subset \mathcal{C}_n$ of π using the following process:

Algorithm 15 (pattern bundle algorithm for $\mathcal{C}_n(\pi)$).

- (1) Let $\pi = \pi_1\pi_2 \cdots \pi_n \in \mathfrak{S}_n$ and write it as a mirrored permutation

$$\pi_1\pi_2 \cdots \pi_n \mid \pi_{n+1}\pi_{n+2} \cdots \pi_{2n} \in \mathfrak{S}_{2n}.$$

- (2) Let $I_n = \{\pi_1, \pi_2, \dots, \pi_n\}$. Fix $0 \leq j \leq n$ and select j elements from the set I_n . Then let Π be the set consisting of the j selected elements.
- (3) The set $I_n \setminus \Pi$ consists of $n - j$ elements. Denote this subset of I_n by Π^c .
- (4) Let $\overline{\Pi^c}$ denote the set containing $\pi_{2n-i_k+1} = 2n - \pi_{i_k} + 1$ for each $\pi_{i_k} \in \Pi^c$. Note that $|\overline{\Pi^c}| = n - j$.
- (5) List the n elements of the set

$$\overline{\Pi^c} \sqcup \Pi$$

so that they are in the same relative order as π and call them $\tau_1\tau_2 \cdots \tau_n$. (Note that the set $\overline{\Pi^c}$ consists of the integers that were switched in Step (4), and the set Π consists of the ones that were fixed in Step (2).)

- (6) The order of the remaining entries $\tau_{n+1}\tau_{n+2} \cdots \tau_{2n}$ is determined by that of $\tau_1\tau_2 \cdots \tau_n$ since we must have $\tau_{2n-i+1} = 2n - \tau_i + 1$ for $1 \leq i \leq n$.
- (7) Output the mirrored permutation $\tau_1\tau_2 \cdots \tau_n \mid \tau_{n+1}\tau_{n+2} \cdots \tau_{2n} \in \mathcal{C}_n \subset \mathfrak{S}_{2n}$ and stop.

Step (5) ensures all of the constructed elements will have the same relative order as π . It follows that the set $\mathcal{C}_n(\pi)$ described in Definition 14 denotes all elements of \mathcal{C}_n created from π by Algorithm 15. Notice in Step (2), we must choose j values to fix. When we let j range from 0 to n , we see that the total number of elements in $\mathcal{C}_n(\pi)$ is given by

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n-1} + \binom{n}{n} = 2^n.$$

We conclude that $|\mathcal{C}_n(\pi)| = 2^n$ for all $\pi \in \mathfrak{S}_n$.

Note that if $\tau = \tau_1 \tau_2 \cdots \tau_n \mid \tau_{n+1} \tau_{n+2} \cdots \tau_{2n} \in \mathcal{C}_n$, then $\tau_1 \cdots \tau_n$ has the same relative order as exactly one element $\pi \in \mathfrak{S}_n$. It follows that if σ and π are distinct permutations of \mathfrak{S}_n , then $\mathcal{C}_n(\sigma) \cap \mathcal{C}_n(\pi) = \emptyset$. Therefore, this process creates all $2^n n!$ elements of \mathcal{C}_n .

Example 16. Using Algorithm 15, we have partitioned the elements of \mathcal{C}_3 into the pattern bundles $\mathcal{C}_n(\pi)$:

$$\mathcal{C}_3 = \left\{ \begin{array}{l} \mathbf{123|456}, \mathbf{132|546}, \mathbf{213|465}, \mathbf{231|645}, \mathbf{312|564}, \mathbf{321|654}, \\ 124|356, 142|536, 214|365, 241|635, 412|563, 421|653, \\ 135|246, 153|426, 315|264, 351|624, 513|462, 531|642, \\ 145|236, 154|326, 326|154, 362|514, 514|362, 541|632, \\ 236|145, 263|415, 415|326, 451|623, 623|451, 632|541, \\ 246|135, 264|315, 426|153, 462|513, 624|351, 642|531, \\ 356|124, 365|214, 536|142, 563|412, 635|241, 653|421, \\ 456|123, 465|213, 546|132, 564|312, 645|231, 654|321 \end{array} \right\}.$$

One can see that the elements of $\pi \in \mathfrak{S}_3$ correspond to the elements in the top row (in bold font). Each column consists of the pattern bundle $\mathcal{C}(\pi)$ corresponding to each $\pi \in \mathfrak{S}_3$.

Definition 17. Let $\pi = \pi_1 \pi_2 \cdots \pi_n \in \mathfrak{S}_n$. We define the *pattern bundle* $\mathcal{D}_n(\pi)$ to be the set of all mirrored permutations $\tau = \tau_1 \tau_2 \cdots \tau_n \mid \tau_{n+1} \cdots \tau_{2n} \in \mathcal{D}_n$ such that $\tau_1 \tau_2 \cdots \tau_n$ has the same relative order as $\pi_1 \pi_2 \cdots \pi_n$.

For every $\pi \in \mathfrak{S}_n$, we construct the subsets $\mathcal{D}_n(\pi) \subset \mathcal{D}_n$ using the following process:

Algorithm 18 (pattern bundle algorithm for $\mathcal{D}_n(\pi)$).

- (1) Let $\pi = \pi_1 \pi_2 \cdots \pi_n \in \mathfrak{S}_n$ and write it as a mirrored permutation

$$\pi_1 \pi_2 \cdots \pi_n \mid \pi_{n+1} \pi_{n+2} \cdots \pi_{2n} \in \mathfrak{S}_{2n}.$$

- (2) If n is even, then pick an even number $2j$, with $0 \leq j \leq \frac{1}{2}n$. Select a subset of $2j$ elements from the set $\{\pi_1, \pi_2, \dots, \pi_n\}$ to keep fixed. Then let Π be the set consisting of the $2j$ selected elements.

If n is odd, then pick an odd number $2j + 1$ with $1 \leq j \leq \frac{1}{2}(n - 1)$. Select a subset of $2j + 1$ elements from the set $\{\pi_1, \pi_2, \dots, \pi_n\}$ to keep fixed. Then let Π be the set consisting of the $2j + 1$ selected elements.

- (3) If n is even, let the set of remaining $n - 2j$ elements be denoted as Π^c . (Note that $n - 2j$ is an even integer.)

If n is odd, let the set of remaining $n - 2j - 1$ elements be denoted as Π^c . (Note that $n - 2j - 1$ is an even integer.)

- (4) Let the set $\overline{\Pi}^c$ denote the set of mirror images from the elements of Π^c . In other words, for each $\pi_{i_k} \in \Pi^c$, the mirror image π_{2n-i_k+1} is in $\overline{\Pi}^c$.
- (5) List elements of the set $\Pi \sqcup \overline{\Pi}^c$ so they are in the same relative order as π and call the resulting permutation $\tau_1 \tau_2 \cdots \tau_n$.
- (6) Then the entries of $\tau_{n+1} \tau_{n+2} \cdots \tau_{2n}$ are determined by the relation $\tau_{2n-i_k+1} = 2n - \tau_{i_k} + 1$.
- (7) Output the mirrored permutation $\tau_1 \tau_2 \cdots \tau_n \mid \tau_{n+1} \tau_{n+2} \cdots \tau_{2n} \in \mathcal{D}_n \subset \mathfrak{S}_{2n}$ and stop.

By Definition 17 the set $\mathcal{D}_n(\pi)$ is the subset of all elements of \mathcal{D}_n which are created from π by Algorithm 18. This is because Step (5) ensures that all of the constructed elements will have the same relative order as π .

In Step (2) we choose an even/odd number of entries to fix, so that we always exchange an even number of entries with their mirror image. This ensures each τ constructed via Algorithm 18 is a type- \mathcal{D}_n mirrored permutation. When n is even, we can see from Step (2) that the total number of permutations created by Algorithm 18 is given by $\sum_{j=0}^{n/2} \binom{n}{2j}$. When n is odd, we can use the identity

$$\sum_{j=0}^{\lfloor n/2 \rfloor} \binom{n}{2j+1} = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n}{2k}, \quad \text{where } 2k = n - (2j + 1),$$

to see that the total number of elements created by Algorithm 18 is also given by the formula

$$\sum_{j=0}^{\lfloor n/2 \rfloor} \binom{n}{2j}.$$

Pascal's identity for computing binomial coefficients states that for all integers n and k with $1 \leq k \leq n - 1$,

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Using this identity we can see that

$$\sum_{j=0}^{\lfloor n/2 \rfloor} \binom{n}{2j} = \sum_{j=0}^{n-1} \binom{n-1}{j} = 2^{n-1}.$$

So for every element $\pi \in \mathfrak{S}_n$, we create 2^{n-1} elements of \mathcal{D}_n . Hence $|\mathcal{D}_n(\pi)| = 2^{n-1}$. Also notice that for each choice of π , the 2^{n-1} elements of $\mathcal{D}_n(\pi)$ will be distinct due to the choice of which elements get sent to their mirror image. Namely, if σ and π are distinct permutations of \mathfrak{S}_n , then $\mathcal{D}_n(\sigma) \cap \mathcal{D}_n(\pi) = \emptyset$. Therefore, this process creates all $2^{n-1}n!$ distinct elements of \mathcal{D}_n .

Example 19. Using Algorithm 18, we have partitioned the set \mathcal{D}_3 into the pattern bundles $\mathcal{D}_n(\pi)$:

$$\mathcal{D}_3 = \left\{ \begin{array}{l} \mathbf{123|456}, \mathbf{132|546}, \mathbf{213|465}, \mathbf{231|645}, \mathbf{312|564}, \mathbf{321|654}, \\ 145|236, 154|326, 214|365, 241|635, 412|563, 421|653, \\ 246|135, 264|315, 426|153, 462|513, 624|351, 642|531, \\ 356|124, 365|214, 536|142, 563|412, 635|241, 653|421 \end{array} \right\}.$$

Note that the elements in the top row (in bold font) are the elements of \mathfrak{S}_3 , while the elements in each column are the elements of the pattern bundle $\mathcal{D}_3(\pi)$ for each respective $\pi \in \mathfrak{S}_3$.

3B. Peak sets in types B and C. Castro-Velez et al. [2013] studied the sets of type- \mathcal{B}_n signed permutations (defined below) with a given peak set $R \subset [n - 1]$. It is well known that the group of signed permutations of type \mathcal{B}_n is isomorphic to the Coxeter groups of types B_n and C_n . In this section, we describe one such isomorphism between the group of signed permutations \mathcal{B}_n and the mirrored permutations \mathcal{C}_n and show how the peak sets in mirrored permutations studied in this paper correspond with the ones studied by Castro-Velez et al. [2013]. It is important to know that even though we compute the cardinalities of similar sets, our methods are completely different and yield different equations. In particular, Castro-Velez et al. use induction arguments similar to those used by Billey, Burdzy, and Sagan in the realm of signed permutations to derive their formulas, whereas we use the pattern bundles of type \mathcal{C}_n to reduce the problem to calculations in the symmetric group.

Let \mathcal{B}_n denote the group of signed permutations on n letters

$$\mathcal{B}_n := \left\{ \beta_1\beta_2 \cdots \beta_n \mid \beta_i \in \{-n, -n + 1, \dots, -1, 1, \dots, n\} \right. \\ \left. \text{and } \{|\beta_1|, |\beta_2|, \dots, |\beta_n|\} = [n] \right\}.$$

We say that a signed permutation $\beta \in \mathcal{B}_n$ has a peak at i if $\beta_{i-1} < \beta_i > \beta_{i+1}$.

Definition 20. Let $R \subseteq [n - 1]$. Then the sets $P_B(R; n)$ and $\hat{P}_B(R; n)$ are defined as

$$P_B(R; n) := \{ \beta \in \mathcal{B}_n \mid P(\beta_1 \cdots \beta_n) = R \}, \\ \hat{P}_B(R; n) := \{ \beta \in \mathcal{B}_n \mid P(\beta_0\beta_1 \cdots \beta_n) = R, \text{ where } \beta_0 = 0 \}.$$

In this paper we study the sets of mirrored permutations of types \mathcal{C}_n and \mathcal{D}_n that have a given peak set S .

Definition 21. Let \mathcal{C}_n and \mathcal{D}_n be the mirrored permutations of types C and D , respectively. For $S \subseteq [n - 1]$, we define the sets $P_C(S; n)$ and $P_D(S; n)$ as

$$P_C(S; n) := \{ \pi \in \mathcal{C}_n \mid P(\pi_1\pi_2 \cdots \pi_n) = S \}, \tag{9}$$

$$P_D(S; n) := \{ \pi \in \mathcal{D}_n \mid P(\pi_1\pi_2 \cdots \pi_n) = S \}. \tag{10}$$

Let $S \subseteq [n]$ we define the sets $\hat{P}_C(S; n)$ and $\hat{P}_D(S; n)$ as

$$\hat{P}_C(S; n) := \{\pi \in \mathcal{C}_n \mid P(\pi_1\pi_2 \cdots \pi_n \mid \pi_{n+1}) = S\}, \tag{11}$$

$$\hat{P}_D(S; n) := \{\pi \in \mathcal{D}_n \mid P(\pi_1\pi_2 \cdots \pi_n \mid \pi_{n+1}) = S\}. \tag{12}$$

Note that $\hat{P}_C(S; n)$ and $\hat{P}_D(S; n)$ differ from $P_C(S; n)$ and $P_D(S; n)$ in that they allow for a peak in the n -th position when $\pi_{n-1} < \pi_n > \pi_{n+1}$. The following proposition provides a bijection between the peak sets $\hat{P}_B(R; n)$ considered by Castro-Velez et al. [2013] and $\hat{P}_C(S; n)$ considered in this paper.

Proposition 22. *Let $S = \{i_1, i_2, \dots, i_k\} \subset \{2, 3, \dots, n\}$ and*

$$R = \{n - i_1 + 1, n - i_2 + 1, \dots, n - i_k + 1\} \subset [n - 1].$$

Then there is a bijection between \mathcal{C}_n and \mathcal{B}_n that maps $\hat{P}_C(S; n)$ to $\hat{P}_B(R; n)$.

The above result states that the peaks of $\pi_1\pi_2 \cdots \pi_n \mid \pi_{n+1}$ correspond bijectively with the peaks of a signed permutation $\beta_0\beta_1\beta_2 \cdots \beta_n$, where $\beta_0 = 0$, and the peaks of $\pi_1\pi_2 \cdots \pi_n$ correspond with those of $\beta_1\beta_2 \cdots \beta_n$. Before proceeding to the proof of Proposition 22, we set some preliminaries.

Billey and Lakshmibai [2000, Definition 8.3.2] note that a mirrored permutation

$$\pi_1\pi_2 \cdots \pi_n \mid \pi_{n+1}\pi_{n+2} \cdots \pi_{2n} \in \mathcal{C}_n$$

can be represented by either side of the mirror, $\pi_1\pi_2 \cdots \pi_n$ or $\pi_{n+1}\pi_{n+2} \cdots \pi_{2n}$, and we use the latter $\pi_{n+1}\pi_{n+2} \cdots \pi_{2n}$ to define a map F from \mathcal{C}_n to \mathcal{B}_n as

$$F : \mathcal{C}_n \rightarrow \mathcal{B}_n,$$

$$\pi_1\pi_2 \cdots \pi_n \mid \pi_{n+1}\pi_{n+2} \cdots \pi_{2n} \mapsto \beta_1\beta_2 \cdots \beta_n,$$

where

$$\beta_i = \begin{cases} \pi_{n+i} - n & \text{if } \pi_{n+i} > n, \\ \pi_{n+i} - n - 1 & \text{if } \pi_{n+i} \leq n. \end{cases}$$

We consider a signed permutation $\beta = \beta_1\beta_2 \cdots \beta_n$ in \mathcal{B}_n as $\beta_0\beta_1 \cdots \beta_n$, where $\beta_0 = 0$, thus allowing for a peak at position 1. We note that the map F respects the relative order of the sequence $\pi_n\pi_{n+1}\pi_{n+2} \cdots \pi_{2n}$; i.e., for $0 \leq i \leq n - 1$, if $\pi_{n+i} < \pi_{n+i+1}$ then $\beta_i < \beta_{i+1}$, and similarly if $\pi_{n+i} > \pi_{n+i+1}$ then $\beta_i > \beta_{i+1}$.

We also define an automorphism $G : \mathcal{B}_n \rightarrow \mathcal{B}_n$ which switches the sign of each β_i in $\beta_0\beta_1\beta_2 \cdots \beta_n$ (keeping $\beta_0 = 0$ fixed). To avoid cumbersome notation, for each β_i , we set $\bar{\beta}_i = -\beta_i$. The following table illustrates how the maps F and G map the group of mirrored permutations \mathcal{C}_2 bijectively to the group of signed permutations \mathcal{B}_2 :

$\pi \in \mathcal{C}_2$	$F(\pi) \in \mathcal{B}_2$	$G(F(\pi)) \in \mathcal{B}_2$
12 34	012	0 $\bar{1}\bar{2}$
21 43	021	0 $\bar{2}\bar{1}$
13 24	0 $\bar{1}\bar{2}$	01 $\bar{2}$
24 13	0 $\bar{2}\bar{1}$	02 $\bar{1}$
31 42	02 $\bar{1}$	0 $\bar{2}\bar{1}$
42 31	01 $\bar{2}$	0 $\bar{1}\bar{2}$
34 12	0 $\bar{2}\bar{1}$	021
43 21	0 $\bar{1}\bar{2}$	012

With the above notation at hand we now proceed to the proof.

Proof of Proposition 22. Let $\pi = \pi_1\pi_2 \cdots \pi_n | \pi_{n+1} \cdots \pi_{2n}$ be a mirrored permutation and $F(\pi) = \beta_0\beta_1\beta_2 \cdots \beta_n$. Then we see $G(F(\pi)) = \beta_0\bar{\beta}_1\bar{\beta}_2 \cdots \bar{\beta}_n$. Suppose $\pi_i < \pi_{i+1}$ for some $i \in \{1, 2, \dots, n\}$. Looking at the mirror images of π_i and π_{i+1} , we get $2n - \pi_i + 1 > 2n - \pi_{i+1} + 1$; thus $\pi_{2n-i+1} > \pi_{2n-(i+1)+1}$. Since the map F respects the relative order of $\pi_n\pi_{n+1} \cdots \pi_{2n}$, we have $\beta_{n-i+1} > \beta_{n-(i+1)+1}$, and thus $\bar{\beta}_{n-i+1} < \bar{\beta}_{n-(i+1)+1}$. Using the same argument but replacing “ $<$ ” with “ $>$ ” and vice versa, we get that if $\pi_i > \pi_{i+1}$ then $\bar{\beta}_{n-i+1} > \bar{\beta}_{n-(i+1)+1}$. Therefore if $\pi \in \hat{P}_C(S; n)$ then $G(F(\pi)) \in \hat{P}_B(R; n)$, and if $\pi \notin \hat{P}_C(S; n)$ then $G(F(\pi)) \notin \hat{P}_B(R; n)$. Since both G and F are bijections, we conclude that $G(F(\hat{P}_C(S; n))) = \hat{P}_B(R; n)$. \square

We can also consider signed permutations $\beta \in \mathcal{B}_n$ without the convention that $\beta_0 = 0$. In that case we obtain the following result.

Corollary 23. *Let $S = \{i_1, i_2, \dots, i_k\} \subset \{2, 3, \dots, n - 1\}$ and*

$$R = \{n - i_1 + 1, n - i_2 + 1, \dots, n - i_k + 1\} \subset \{2, \dots, n - 1\}.$$

Then the bijection $G \circ F : \mathcal{C}_n \rightarrow \mathcal{B}_n$ maps $P_C(S; n)$ to $P_B(R; n)$.

Proof. The proof of this corollary proceeds exactly as the proof of Proposition 22. \square

3C. The sets $P_C(S; n)$ and $P_D(S; n)$. In this subsection, we use the fact that $\mathcal{C}_n(\pi)$ and $\mathcal{D}_n(\pi)$ partition \mathcal{C}_n and \mathcal{D}_n to give concise proofs that $|P_C(S; n)| = p(n)2^{2n-|S|-1}$ and $|P_D(S; n)| = p(n)2^{2n-|S|-2}$, where $p(n)$ is the polynomial given in [Billey et al. 2013, Theorem 2.2].

Theorem 24. *Let $S \subseteq [n - 1]$. Then*

- (I) $|P_C(S; n)| = p(n)2^{2n-|S|-1}$,
- (II) $|P_D(S; n)| = p(n)2^{2n-|S|-2}$.

Proof. To prove part (I), note that Billey et al. [2013, Theorem 2.2] showed that $|P(S; n)| = p(n)2^{n-|S|-1}$, where $p(n)$ is a polynomial with degree $\max(S) - 1$. Algorithm 15 showed that each $\pi \in P(S; n)$ corresponds to a subset $\mathcal{C}_n(\pi) \subset \mathcal{C}_n$

which contains 2^n elements. By construction these elements have the exact same peak set as π . In other words, for every $\tau \in \mathcal{C}_n(\pi)$, the peak sets $P(\tau) = P(\pi) = S$ agree. We compute that $|P_C(S; n)| = p(n)2^{n-|S|-1}2^n = p(n)2^{2n-|S|-1}$.

Part (II) follows similarly, replacing \mathcal{C}_n with \mathcal{D}_n , Algorithm 15 with Algorithm 18, and 2^n with 2^{n-1} . □

4. Peak sets of the Coxeter groups of types C and D

In this section we use specific sums of binomial coefficients and the partitions

$$P(S; n) = \overline{P(S; n)} \sqcup \underline{P(S; n)} \quad \text{and} \quad \overline{P(S; n)} = \bigsqcup_{k=2}^n P(S; n)^{\nearrow k}$$

to describe the cardinality of the sets $\hat{P}_C(S; n)$, $\hat{P}_C(S \cup \{n\}; n)$, $\hat{P}_D(S; n)$, and $\hat{P}_D(S \cup \{n\}; n)$. We begin by setting the following notation:

Definition 25. Let $\Phi(n, k)$ denote the sum of the last $n - j + 1$ terms of the n -th row in Pascal’s triangle,

$$\Phi(n, k) = \sum_{i=k}^n \binom{n}{i} = \binom{n}{k} + \binom{n}{k+1} + \cdots + \binom{n}{n},$$

and let

$$\Psi(n, k) = 2^n - \Phi(n, k).$$

We can now state our main result.

Theorem 26. *Type C:* Let $\hat{P}_C(S; n)$ denote the set of elements of \mathcal{C}_n with peak set $S \subset [n - 1]$. Then

$$|\hat{P}_C(S; n)| = \sum_{k=1}^n |P(S; n)^{\nearrow k}| \cdot \Phi(n, k) + |\underline{P(S; n)}| \cdot 2^n$$

and

$$|\hat{P}_C(S \cup \{n\}; n)| = \sum_{k=1}^n |P(S; n)^{\nearrow k}| \cdot \Psi(n, k).$$

Type D: Let $\hat{P}_D(S; n)$ denote the set of elements of \mathcal{D}_n with peak set $S \subset [n - 1]$. If n is even, then

$$|\hat{P}_D(S; n)| = \sum_{k=1}^{n/2} (|P(S; n)^{\nearrow 2k-1}| + |P(S; n)^{\nearrow 2k}|) \Phi(n - 1, 2k - 1) + |\underline{P(S; n)}| 2^{n-1}$$

and

$$|\hat{P}_D(S \cup \{n\}; n)| = \sum_{k=1}^{n/2} (|P(S; n)^{\nearrow 2k-1}| + |P(S; n)^{\nearrow 2k}|) \Psi(n - 1, 2k - 1).$$

If n is odd, then

$$|\hat{P}_D(S; n)| = \sum_{k=1}^{(n-1)/2} (|P(S; n)^{\nearrow 2k+1}| + |P(S; n)^{\nearrow 2k}|) \Phi(n-1, 2k) + |P(S; n)| 2^{n-1}$$

and

$$|\hat{P}_D(S \cup \{n\}; n)| = \sum_{k=1}^{(n-1)/2} (|P(S; n)^{\nearrow 2k+1}| + |P(S; n)^{\nearrow 2k}|) \Psi(n-1, 2k).$$

Since the proofs of the type- C and type- D results in Theorem 26 require some specific identities involving the functions Φ and Ψ , we present these results and proofs in Sections 4A and 4B, respectively.

Note that Proposition 22 shows that $|\hat{P}_B(R; n)| = |\hat{P}_C(S; n)|$. Castro-Velez et al. [2013, Theorem 3.2] gave a recursive formula for computing the cardinality of the set $\hat{P}_B(R; n)$. Theorem 26 provides an alternate formula for $|\hat{P}_C(S; n)| = |\hat{P}_B(R; n)|$ using the sums of binomial coefficients $\Phi(n, k)$ and $\Psi(n, k)$, and the cardinalities of sets $\underline{P}(S; n)$ and $P(S; n)^{\nearrow k}$.

4A. Peak sets of the Coxeter groups of type C . The following lemma uses the functions $\Phi(n, k)$ and $\Psi(n, k)$ to count the number of elements in $\mathcal{C}_n(\pi)$ having an ascent in the n -th position. This lemma is the key step in the type- C proof of Theorem 26.

Lemma 27. *If $\pi \in P(S; n)^{\nearrow k}$ then there are $\Phi(n, k)$ elements $\tau \in \mathcal{C}_n(\pi)$ with $\tau_n \leq n$ and $\Psi(n, k)$ elements $\tau \in \mathcal{C}_n(\pi)$ with $\tau_n > n$.*

Proof. Suppose that $\pi = \pi_1\pi_2 \cdots \pi_n \in P(S; n)^{\nearrow k}$, so $\pi_{n-1} < \pi_n = k$. If $\tau = \tau_1\tau_2 \cdots \tau_n \mid \tau_{n+1}\tau_{n+2} \cdots \tau_{2n} \in \mathcal{C}_n(\pi)$, then τ_n is the k -th largest integer in the set $\{\tau_1, \tau_2, \dots, \tau_n\}$ because τ has the same relative order as π and $\pi_n = k$. Therefore if at least k elements of the set $\{\tau_1, \tau_2, \dots, \tau_n\}$ have $\tau_i \leq n$ then we conclude $\tau_n \leq n$.

We will show there are $\binom{n}{j}$ elements $\tau \in \mathcal{C}_n(\pi)$, where exactly j elements of the set $\{\tau_1, \tau_2, \dots, \tau_n\}$ satisfy $\tau_n \leq n$. To construct such a τ , we start with $\pi = \pi_1\pi_2 \cdots \pi_n$, and then we choose j elements of the set $\{\pi_1, \pi_2, \dots, \pi_n\}$ to remain fixed. We replace the remaining $n - j$ elements of $\{\pi_1, \pi_2, \dots, \pi_n\}$ with their mirror images, which are all greater than n . Finally, we list the elements of the resulting set so that they have the same relative order as π and call them $\tau_1\tau_2 \cdots \tau_n$. The subpermutation $\tau_{n+1}\tau_{n+2} \cdots \tau_{2n}$ is then completely determined by the subpermutation $\tau_1\tau_2 \cdots \tau_n$. Thus there are $\binom{n}{j}$ mirrored permutations τ of the form $\tau = \tau_1\tau_2 \cdots \tau_n \mid \tau_{n+1}\tau_{n+2} \cdots \tau_{2n} \in \mathcal{C}_n(\pi)$, where j of the elements in $\{\tau_1, \tau_2, \dots, \tau_n\}$ satisfy $\tau_i \leq n$.

Considering all integers j with $k \leq j \leq n$, we see that the number of elements

$$\tau = \tau_1\tau_2 \cdots \tau_n \mid \tau_{n+1}\tau_{n+2} \cdots \tau_{2n} \in \mathcal{C}_n(\pi)$$

with at least k of the elements in $\{\tau_1, \tau_2, \dots, \tau_n\}$ satisfying $\tau_i \leq n$ is exactly

$$\Phi(n, k) = \sum_{j=k}^n \binom{n}{j}.$$

Thus there are $\Phi(n, k)$ elements $\tau \in \mathcal{C}_n(\pi)$ with $\tau_n \leq n$. The other $2^n - \Phi(n, k) = \Psi(n, k)$ elements of $\mathcal{C}_n(\pi)$ must have $\tau_n > n$. □

With the above result at hand, we now give the following proof.

Proof of Theorem 26, type C. Let $\pi = \pi_1\pi_2 \cdots \pi_n \in P(S; n)$, and recall that $\mathcal{C}_n(\pi)$ is the set of elements of \mathcal{C}_n whose first n entries have the same relative order as π , and $|\mathcal{C}_n(\pi)| = 2^n$ for any $\pi \in \mathfrak{S}_n$. Let $\tau = \tau_1\tau_2 \cdots \tau_n \mid \tau_{n+1} \cdots \tau_{2n} \in \mathcal{C}_n$ denote a mirrored permutation of type \mathcal{C}_n . Then there are two possibilities:

- Either τ has the same peak set as π so that $\tau \in \hat{P}_C(S; n)$, or
- τ has an additional peak at n , in which case $\tau \in \hat{P}_C(S \cup \{n\}; n)$.

There are two cases in which $\tau \in \hat{P}_C(S; n)$:

Case 1: If π ends with a descent, i.e., $\pi_{n-1} > \pi_n$, then every $\tau \in \mathcal{C}_n(\pi)$ also has $\tau_{n-1} > \tau_n$, and thus τ is in $\hat{P}_C(S; n)$ because it cannot possibly have a peak at n if it ends with a descent. We conclude that if $\pi \in \underline{P(S; n)}$ then all 2^n elements $\tau \in \mathcal{C}_n(\pi)$ are in $\hat{P}_C(S; n)$.

Case 2: If π ends with an ascent, i.e., $\pi_{n-1} < \pi_n$, then $\tau_{n-1} < \tau_n$ for all $\tau \in \mathcal{C}_n(\pi)$ as well. (Recall that for any $\sigma \in \mathcal{C}_n$, our map into \mathfrak{S}_{2n} identifies σ_i with its *mirror* σ_{n-i+1} by $\sigma_{n-i+1} = 2n - \sigma_i + 1$.) Hence, if $\tau_n \leq n$, then $\tau_{n+1} = 2n - \tau_n + 1 > \tau_n$. In this case $\tau_{n-1} < \tau_n < \tau_{n+1}$, and τ does not have a peak at n . So $\tau \in \hat{P}_C(S; n)$. Therefore we conclude that if $\pi \in \overline{P(S; n)}$ and if $\tau \in \mathcal{C}_n(\pi)$ satisfies $\tau_n \leq n$ then τ is an element of $\hat{P}_C(S; n)$. By Lemma 27 we conclude that if $\pi \in P(S; n)^{\nearrow k}$ then $\Phi(n, k)$ of the elements in $\mathcal{C}_n(\pi)$ are in $\hat{P}_C(S; n)$.

Case 3: There is only one case in which $\tau \in \hat{P}_C(S \cup \{n\}; n)$. If $\pi \in \overline{P(S; n)}$ and $\tau \in \mathcal{C}_n(\pi)$ is such that $\tau_n > n$, then τ must satisfy $\tau_{n-1} < \tau_n > \tau_{n+1}$ because $\tau_{n+1} = 2n - \tau_n + 1 < n$. Therefore τ is an element of $\hat{P}_C(S \cup \{n\}; n)$. Applying Lemma 27 we conclude that if $\pi \in P(S; n)^{\nearrow k}$ then $\Psi(n, k)$ of the elements in $\mathcal{C}_n(\pi)$ are in $\hat{P}_C(S \cup \{n\}; n)$.

From Cases 1 and 2, we conclude that the cardinality of $\hat{P}_C(S; n)$ is given by

$$|\hat{P}_C(S; n)| = \sum_{k=1}^n |P(S; n)^{\nearrow k}| \cdot \Phi(n, k) + |\underline{P(S; n)}| \cdot 2^n.$$

From Case 3, we get

$$|\hat{P}_C(S \cup \{n\}; n)| = \sum_{k=1}^n |P(S; n)^{\nearrow k}| \cdot \Psi(n, k). \quad \square$$

The following example illustrates the type- C formulas proven in Theorem 26.

Example 28. Using the results of this section, we compute the sets $\hat{P}_C(S; 3)$, where $S \subset [3]$. First, the group \mathfrak{S}_3 can be partitioned as $\mathfrak{S}_3 = P(\emptyset; 3) \sqcup P(\{2\}; 3)$, where

$$P(\emptyset; 3) = \{123, 321, 213, 312\} \quad \text{and} \quad P(\{2\}; 3) = \{132, 231\}.$$

To calculate the peak sets in type C_3 , we will further partition the sets $P(\emptyset; 3)$ and $P(\{2\}; 3)$ using Definition 1. Hence we compute

$$P(\emptyset; 3) = \underline{P(\emptyset; 3)} \sqcup P(\emptyset; 3)^{\nearrow 2} \sqcup P(\emptyset; 3)^{\nearrow 3},$$

where $\underline{P(\emptyset; 3)} = \{321\}$, $P(\emptyset; 3)^{\nearrow 2} = \{312\}$, and $P(\emptyset; 3)^{\nearrow 3} = \{123, 213\}$.

We also compute the set

$$P(\{2\}; 3) = \underline{P(\{2\}; 3)} = \{231, 132\}.$$

Of the 48 elements of the Coxeter group C_3 , only $2^3|P(\emptyset; 3)| = 2^3 \cdot 4 = 32$ elements are in $\hat{P}_C(\emptyset; 3) \sqcup \hat{P}_C(\{3\}; 3)$. Of these 32 permutations, we observe that 18 lie in $\hat{P}_C(\{3\}; 3)$ and 14 lie in $\hat{P}_C(\emptyset; 3)$. We calculate $|\hat{P}_C(\emptyset; 3)|$ using Theorem 26:

$$\begin{aligned} |\hat{P}_C(\emptyset; 3)| &= (|\underline{P(\emptyset; 3)}| \cdot 2^3) + (|P(\emptyset; 3)^{\nearrow 2}| \cdot \Phi(3, 2)) + (|P(\emptyset; 3)^{\nearrow 3}| \cdot \Phi(3, 3)) \\ &= (1 \cdot 8) + (1 \cdot 4) + (2 \cdot 1) = 14. \end{aligned}$$

Hence $|\hat{P}_C(\{3\}; 3)| = 2^3 \cdot 4 - 14 = 18$. Since $P(\{2\}; 3) = \underline{P(\{2\}; 3)}$, we have

$$|\hat{P}_C(\{2\}; 3)| = |\underline{P(\{2\}; 3)}| \cdot 2^3 = |P(\{2\}; 3)| \cdot 2^3 = 16.$$

Indeed one may confirm that $\hat{P}_C(\{2\}; 3)$ is the union of the sets

$$C_3(231) = \left\{ \begin{array}{l} 231|645, \quad 241|635, \\ 351|624, \quad 362|514, \\ 451|623, \quad 462|513, \\ 356|124, \quad 564|312 \end{array} \right\} \quad \text{and} \quad C_3(132) = \left\{ \begin{array}{l} 132|645, \quad 142|536, \\ 153|426, \quad 263|415, \\ 154|326, \quad 264|315, \\ 365|214, \quad 465|213 \end{array} \right\}.$$

4B. Peak sets of the Coxeter group of type D . In this section, we use the functions $\Phi(n, k)$ and $\Psi(n, k)$ to describe the cardinalities of $\hat{P}_D(S; n)$ and $\hat{P}_D(S \cup \{n\}; n)$. The results depend on the parity of n . We begin by providing the following lemmas (similar to Lemma 27), which are used in the type- D proof of Theorem 26.

Lemma 29. *Let n be even, and let $1 \leq k \leq \frac{1}{2}n$. If $\pi \in P(S; n)^{\nearrow 2k} \sqcup P(S; n)^{\nearrow 2k-1}$, then there are $\Phi(n-1, 2k-1)$ elements $\tau \in \mathcal{D}_n(\pi)$ with $\tau_n \leq n$, and $\Psi(n-1, 2k-1)$ elements $\tau \in \mathcal{D}_n(\pi)$ with $\tau_n > n$.*

Proof. Suppose $\pi = \pi_1\pi_2 \cdots \pi_n \in P(S; n)^{\nearrow 2k}$, so $\pi_n = 2k$ and $\pi_{n-1} = i$ for some integer $i < 2k$. If $\tau = \tau_1\tau_2 \cdots \tau_n | \tau_{n+1}\tau_{n+2} \cdots \tau_{2n} \in \mathcal{D}_n(\pi)$, then τ_n is the $2k$ -th largest integer in the set $\{\tau_1, \tau_2, \dots, \tau_n\}$ because τ has the same relative order

as π and $\pi_n = 2k$. Therefore if at least $2k$ elements of the set $\{\tau_1, \tau_2, \dots, \tau_n\}$ satisfy $\tau_i \leq n$ then we can conclude that $\tau_n \leq n$. Moreover, $\tau_n \leq n$ if and only if $\tau_n < \tau_{n+1} = 2n - \tau_n + 1$. Thus we wish to count the number of $\tau \in \mathcal{D}_n(\pi)$ with $\tau_n \leq n$.

In the construction of $\mathcal{D}_n(\pi)$, the total number of τ with at least $2k$ of the elements from $\{\tau_1, \tau_2, \dots, \tau_n\}$ fixed (and less than or equal to n) is given by the sum

$$\binom{n}{2k} + \binom{n}{2k+2} + \dots + \binom{n}{n-2} + \binom{n}{n} \tag{13}$$

when n is even. Using the identity $\binom{n}{2k} = \binom{n-1}{2k-1} + \binom{n-1}{2k}$, we can see that the quantity in (13) equals

$$\left[\binom{n-1}{2k-1} + \binom{n-1}{2k} \right] + \left[\binom{n-1}{2k+1} + \binom{n-1}{2k+2} \right] + \dots + \binom{n-1}{n-1} = \Phi(n-1, 2k-1)$$

when n is even.

Suppose $\pi = \pi_1\pi_2 \dots \pi_n \in P(S; n)^{\nearrow 2k-1}$, so $\pi_n = 2k - 1$ and $\pi_{n-1} = i$ for some integer $i < 2k - 1$. If $\tau = \tau_1\tau_2 \dots \tau_n \mid \tau_{n+1}\tau_{n+2} \dots \tau_{2n} \in \mathcal{D}_n(\pi)$, then τ_n is the $(2k-1)$ -th largest integer in the set $\{\tau_1, \tau_2, \dots, \tau_n\}$ because τ has the same relative order as π and $\pi_n = 2k - 1$. Therefore if at least $2k - 1$ elements of the set $\{\tau_1, \tau_2, \dots, \tau_n\}$ satisfy $\tau_i \leq n$ then we can conclude that $\tau_n \leq n$. Moreover, $\tau_n \leq n$ if and only if $\tau_n < \tau_{n+1} = 2n - \tau_n + 1$. So again, the number of elements with $\tau_n \leq n$ is $\Phi(n - 1, 2k - 1)$.

We conclude that when $\pi \in P(S; n)^{\nearrow 2k} \sqcup P(S; n)^{\nearrow 2k-1}$, there are $\Phi(n-1, 2k-1)$ mirrored permutations $\tau \in \mathcal{D}_n(\pi)$ with $\tau_n < \tau_{n+1}$. Since there are 2^{n-1} elements in $\mathcal{D}_n(\pi)$, we see that there are $\Psi(n-1, 2k-1)$ elements $\tau \in \mathcal{D}_n(\pi)$ with $\tau_n > \tau_{n+1}$. \square

Lemma 30. *Let n be odd and let $1 \leq k \leq \frac{1}{2}(n - 1)$. If $\pi \in P(S; n)^{2k}$ or $\pi \in P(S; n)^{2k+1}$ then there are $\Phi(n - 1, 2k)$ elements $\tau \in \mathcal{D}_n(\pi)$ with $\tau_n \leq n$ and $\Psi(n - 1, 2k)$ elements $\tau \in \mathcal{D}_n(\pi)$ with $\tau_n > n$.*

The proof of Lemma 30 follows similarly to that of Lemma 29; hence we omit the argument, but point the interested reader to the arXiv preprint of this paper for a detailed proof [Diaz-Lopez et al. 2015]. We are now ready to enumerate the sets $\hat{P}_D(S; n)$ and $\hat{P}_D(S \cup \{n\}; n)$.

Proof of Theorem 26, type D. Let $\pi \in P(S; n)$, $\tau = \tau_1\tau_2 \dots \tau_n \mid \tau_{n+1}\tau_{n+2} \dots \tau_{2n} \in \mathcal{D}_n$, and recall that $\mathcal{D}_n(\pi)$ consists of the elements of \mathcal{D}_n which have the same relative order as π . There are 2^{n-1} such elements. Since $\tau \in \mathcal{D}_n(\pi)$, its first n entries $\tau_1\tau_2 \dots \tau_n$ have the same relative order as $\pi_1\pi_2 \dots \pi_n$, and just as in the type- \mathcal{C}_n case, there are two possibilities:

- Either τ has the same peak set as π so that $\tau \in \hat{P}_D(S; n)$, or
- τ has an additional peak at n , in which case $\tau \in \hat{P}_D(S \cup \{n\}; n)$.

There are two cases in which $\tau \in \hat{P}_D(S; n)$:

Case 1: If π ends with a descent, i.e., $\pi_{n-1} > \pi_n$, then every $\tau \in \mathcal{D}_n(\pi)$ also has $\tau_{n-1} > \tau_n$, and thus τ is in $\hat{P}_D(S; n)$ because it cannot possibly have a peak at n if it has a descent at $n - 1$. We conclude that if $\pi \in \overline{P(S; n)}$ then all 2^{n-1} elements of $\mathcal{D}_n(\pi)$ are in $\hat{P}_D(S; n)$.

Case 2: If π ends with an ascent, $\pi_{n-1} < \pi_n$, then $\tau_{n-1} < \tau_n$ for all $\tau \in \mathcal{D}_n(\pi)$ as well. (Recall that for any $\sigma \in \mathcal{D}_n$, our map into \mathfrak{S}_{2n} identifies σ_i with σ_{n-i+1} by $\sigma_{n-i+1} = 2n - \sigma_i + 1$.) Hence, if $\tau_n \leq n$, then $\tau_{n+1} = 2n - \tau_n + 1 > \tau_n$. In this case $\tau_{n-1} < \tau_n < \tau_{n+1}$, and τ does not have a peak at n . So $\tau \in \hat{P}_D(S; n)$. Therefore we conclude that if $\pi \in \overline{P(S; n)}$ and if $\tau \in \mathcal{D}_n(\pi)$ satisfies $\tau_n \leq n$ then τ is an element of $\hat{P}_D(S; n)$. By Lemma 29 we conclude that if $\pi \in P(S; n)^{\nearrow k}$ then $\Phi(n - 1, 2k - 1)$ of the elements in $\mathcal{D}_n(\pi)$ are in $\hat{P}_D(S; n)$.

Case 3: There is only one case in which $\tau \in \hat{P}_D(S \cup \{n\}; n)$. If $\pi \in \overline{P(S; n)}$ and $\tau \in \mathcal{D}_n(\pi)$ is such that $\tau_n > n$, then τ must satisfy $\tau_{n-1} < \tau_n > \tau_{n+1}$ because $\tau_{n+1} = 2n - \tau_n + 1 < n$. Therefore τ is an element of $\hat{P}_D(S \cup \{n\}; n)$.

We have shown if π is in $\overline{P(S; n)}$, then all 2^{n-1} elements $\mathcal{D}_n(\pi)$ are in $\hat{P}_D(S; n)$. Lemma 29 showed when n is even and $\pi \in P(S; n)^{\nearrow 2k}$ or $\pi \in P(S; n)^{\nearrow 2k-1}$, then $\Phi(n, 2k - 1)$ of the elements of $\mathcal{D}_n(\pi)$ are in $\hat{P}_D(S; n)$. Thus we conclude when n is even, the cardinality of $\hat{P}_D(S; n)$ is given by the formula

$$|\hat{P}_D(S; n)| = \sum_{k=1}^n (|P(S; n)^{k-1}| + |P(S; n)^{2k}|) \cdot \Phi(n, 2k - 1) + |\overline{P(S; n)}| \cdot 2^{n-1}.$$

Lemma 29 also showed if $\pi \in P(S; n)^{\nearrow 2k}$ or $\pi \in P(S; n)^{\nearrow 2k-1}$ then $\Psi(n-1, 2k-1)$ elements from $\mathcal{D}_n(\pi)$ are in the set $\hat{P}_D(S \cup \{n\}; n)$, and thus

$$|\hat{P}_D(S \cup \{n\}; n)| = \sum_{k=1}^{n/2} (|P(S; n)^{\nearrow 2k-1}| + |P(S; n)^{\nearrow 2k}|) \cdot \Psi(n - 1, 2k - 1)$$

when n is even.

Lemma 30 showed when n is odd and $\pi \in P(S; n)^{\nearrow 2k}$ or $\pi \in P(S; n)^{\nearrow 2k-1}$, then $\Phi(n - 1, 2k)$ of the elements of $\mathcal{D}_n(\pi)$ are in $\hat{P}_D(S; n)$. Thus we conclude that when n is odd, the cardinality of $\hat{P}_D(S; n)$ is given by the formula

$$|\hat{P}_D(S; n)| = \sum_{k=1}^{(n-1)/2} (|P(S; n)^{\nearrow 2k+1}| + |P(S; n)^{\nearrow 2k}|) \cdot \Phi(n - 1, 2k) + |\overline{P(S; n)}| \cdot 2^{n-1}.$$

Lemma 30 also showed if $\pi \in P(S; n)^{\nearrow 2k}$ or $\pi \in P(S; n)^{\nearrow 2k+1}$ then $\Psi(n - 1, 2k)$ elements from $\mathcal{D}_n(\pi)$ are in the set $\hat{P}_D(S \cup \{n\}; n)$, and thus

$$|\hat{P}_D(S \cup \{n\}; n)| = \sum_{k=1}^{(n-1)/2} (|P(S; n)^{\nearrow 2k+1}| + |P(S; n)^{\nearrow 2k}|) \cdot \Psi(n - 1, 2k)$$

when n is odd. This proves the formula for the cardinality of $\hat{P}_D(S \cup \{n\}; n)$. \square

4C. Special case: empty peak set in types C and D. In this section we consider the special case of $S = \emptyset$ in types C_n and D_n .

Proposition 31. *Let $n \geq 2$ and $m \geq 4$, then*

- (I) $|\hat{P}_C(\emptyset; n)| = \frac{1}{2}(3^n + 1),$
- (II) $|\hat{P}_D(\emptyset; m)| = \frac{1}{4}3^m + \frac{1}{4}(-1)^m + \frac{1}{2}.$

Proposition 31(I) was originally proved by Castro-Velez et al. [2013, Theorem 2.4] in type B_n . However, the proof given here is a combinatorial argument involving ternary sequences (in the letters A, B and C) with an even number of B 's that restricts naturally to a proof of a similar result involving the mirrored permutations with no peaks in type D_n as well.

The integer sequence given by Proposition 31(I) is sequence A007051 in [OEIS 1996] after the first three iterations. Let T_n denote the set of ternary sequences (in the letters A, B and C) of length n with an even number of B 's. It is noted on Sloane's OEIS that $\frac{1}{2}(3^n + 1)$ counts all such sequences.

Proof of Proposition 31(I). To prove that $|\hat{P}_C(\emptyset; n)| = |T_n| = \frac{1}{2}(3^n + 1)$, we prove there is a bijection between the sets T_n and $\hat{P}_C(\emptyset; n)$.

Every permutation $\pi \in \hat{P}_C(\emptyset; n)$ has the form $\pi = \pi_A \pi_B \pi_C | \overline{\pi_C \pi_B \pi_A}$, where π_A is a sequence of numbers in descending order and each $\pi_i \in \pi_A$ is greater than n , π_B is a sequence of numbers in descending order and each $\pi_i \in \pi_B$ is less than or equal to n , and π_C is a sequence of numbers in ascending order and each $\pi_i \in \pi_C$ is less than or equal to n . Note that the mirror image $\overline{\pi_C \pi_B \pi_A}$ is determined uniquely by $\pi_A \pi_B \pi_C$, so to condense notation in this proof we will refrain from writing it. It is possible for at most two of the parts π_A, π_B , or π_C to be empty. Moreover, there is always a choice of whether to include the minimum element of the subpermutation $\pi_B \pi_C$ as the last element in π_B or the first element in π_C . We always choose to make the length of π_B even by including/excluding this minimum element depending on the parity of π_B .

More precisely, let $\pi = \pi_A \pi_B \pi_C \in \hat{P}_C(\emptyset; n)$, where

$$\pi_A = [\pi_1 > \dots > \pi_k], \quad \pi_B = [\pi_{k+1} > \dots > \pi_{k+j}], \quad \text{and} \quad \pi_C = [\pi_{k+j+1} < \dots < \pi_n].$$

Define a set map $\Delta : \hat{P}_C(\emptyset; n) \rightarrow T_n$ by assigning a ternary sequence $\Delta(\pi) = x$ in T_n to each element $\pi \in \hat{P}_C(\emptyset; n)$ by setting

$$\Delta(\pi)_i = x_i = \begin{cases} A & \text{if } i \in \{2n - \pi_1 + 1, \dots, 2n - \pi_k + 1\}, \\ B & \text{if } i \in \{\pi_{k+1}, \dots, \pi_{k+j}\}, \\ C & \text{if } i \in \{\pi_{k+j+1}, \dots, \pi_n\}. \end{cases}$$

Note that there is an even number of B 's by the way we defined π_B . Hence $\Delta(\pi) = x \in T_n$.

We can also define a set map $\Theta : T_n \rightarrow \hat{P}_C(\emptyset; n)$ by reversing this process. That is to say, given a ternary sequence $x = x_1x_2 \cdots x_n$ in T_n , define \mathcal{A}, \mathcal{B} , and \mathcal{C} as

$$\mathcal{A} = \{1 \leq i \leq n : x_i = A\}, \quad \mathcal{B} = \{1 \leq i \leq n : x_i = B\}, \quad \text{and} \quad \mathcal{C} = \{1 \leq i \leq n : x_i = C\}.$$

List the elements of \mathcal{A} and \mathcal{C} in ascending order and \mathcal{B} in descending order:

$$\begin{aligned} \mathcal{A} &= [a_1 < a_2 < \cdots < a_k], & \mathcal{B} &= [b_{k+1} > b_{k+2} > \cdots > b_{k+j}], \\ \mathcal{C} &= [c_{k+j+1} < c_{k+j+2} < \cdots < c_n]. \end{aligned}$$

Then define $\Theta(x) = \pi$, where

$$\pi_i = \begin{cases} 2n - a_i + 1 & \text{if } 1 \leq i \leq k, \\ b_i & \text{if } k + 1 \leq i \leq k + j, \\ c_i & \text{if } k + j + 1 \leq i \leq n. \end{cases}$$

Notice that after π_i is determined for $1 \leq i \leq n$, the rest of π is determined.

To show $\Theta \circ \Delta = \text{Id}$, let $\pi = \pi_A \pi_B \pi_C \in \hat{P}_C(\emptyset; n)$, where

$$\pi_A = [\pi_1 > \cdots > \pi_k], \quad \pi_B = [\pi_{k+1} > \cdots > \pi_{k+j}], \quad \text{and} \quad \pi_C = [\pi_{k+j+1} < \cdots < \pi_n],$$

and set $\sigma = \Theta(\Delta(\pi)) = \sigma_1 \cdots \sigma_n$. Then

$$\Delta(\pi)_i = x_i = A \quad \text{for } i \in \{2n - \pi_1 + 1, \dots, 2n - \pi_k + 1\},$$

so $\mathcal{A} = [2n - \pi_1 + 1 < \cdots < 2n - \pi_k + 1]$. By the definition of Θ , we get $\sigma_i = 2n - (2n - \pi_i + 1) + 1$ for $1 \leq i \leq k$; thus $\sigma_i = \pi_i$ for $1 \leq i \leq k$.

Similarly, $\Delta(\pi)_i = x_i = B$ for $i \in \{\pi_{k+1}, \dots, \pi_{k+j}\}$; thus $\mathcal{B} = [\pi_{k+1} > \cdots > \pi_{k+j}]$. By the definition of Θ , we get $\sigma_i = \pi_i$ for $k + 1 \leq i \leq k + j$. Finally, $\Delta(\pi)_i = x_i = C$ for $i \in \{\pi_{k+j+1}, \dots, \pi_n\}$; thus $\mathcal{C} = [\pi_{k+j+1} < \cdots < \pi_n]$. By the definition of Θ , we see that $\sigma_i = \pi_i$ for $k + j + 1 \leq i \leq n$. Therefore $\sigma_i = \pi_i$ for $1 \leq i \leq n$, which implies $\Theta(\Delta(\pi)) = \sigma = \pi$ for all $\pi \in \hat{P}_C(\emptyset; n)$. A similar argument shows $\Delta(\Theta(x)) = x$ for all $x \in T_n$. □

The integer sequence given by Proposition 31(II) is sequence A122983 in [OEIS 1996] after the first three iterations. To prove this result, we let T_n denote the set of ternary sequences (in the letters A, B and C) of length n with an even number of A 's and B 's. It is noted on Sloane's OEIS that $\frac{1}{4}3^n + \frac{1}{4}(-1)^n + \frac{1}{2}$ counts all such sequences. In the following proof we construct a bijection from T_n to $\hat{P}_D(\emptyset; n)$ by using the maps Δ and Θ , similar to the proof of Proposition 31(I).

Proof of Proposition 31(II). The proof follows as the proof of Proposition 31(I), with the additional condition that the length of π_A is even since every element π in $\hat{P}_D(\emptyset; n)$ has an even number of entries in $\pi_1\pi_2 \cdots \pi_n$ that are greater than n . We point the interested reader to the arXiv preprint version of this paper for a detailed proof [Diaz-Lopez et al. 2015]. □

We will illustrate the bijection between Δ and Θ , described in the proof of Proposition 31, with the following example.

Example 32. Type C: Consider the permutation $\pi \in \mathcal{C}_{10}$, where

$$\pi = 20\ 18\ 13\ 10\ 9\ 7\ 4\ 2\ 5\ 6 \mid 15\ 16\ 19\ 17\ 14\ 12\ 11\ 8\ 3\ 1.$$

Let $\Delta(\pi) = x \in T_n$. Since

$$\pi_A = 20\ 18\ 13, \quad \pi_B = 10\ 9\ 7\ 4, \quad \text{and} \quad \pi_C = 2\ 5\ 6,$$

we have $x_i = A$ for $i \in \{1, 3, 8\}$, $x_i = B$ for $i \in \{4, 7, 9, 10\}$, and $x_i = C$ for $i \in \{2, 5, 6\}$. Thus $\Delta(\pi) = x = ACABCCBABB$.

Consider $\Theta(\Delta(\pi)) \in \hat{P}_C(\emptyset; 10)$. Since $\Delta(\pi) = x = ACABCCBABB$, the lists \mathcal{A} , \mathcal{B} and \mathcal{C} are defined as

$$\mathcal{A} = [1 < 3 < 8], \quad \mathcal{B} = [10 > 9 > 7 > 4], \quad \text{and} \quad \mathcal{C} = [2 < 5 < 6].$$

Using the definition of Θ , we get

$$\Theta(\Delta(\pi)) = \Theta(x) = 20\ 18\ 13\ 10\ 9\ 7\ 4\ 2\ 5\ 6 \mid 15\ 16\ 19\ 17\ 14\ 12\ 11\ 8\ 3\ 1 = \pi.$$

Type D: Consider the permutation $\pi \in \mathcal{D}_{10}$, where

$$\pi = 20\ 18\ 13\ 11\ 9\ 7\ 4\ 2\ 5\ 6 \mid 15\ 16\ 19\ 17\ 14\ 12\ 10\ 8\ 3\ 1.$$

Let $\Delta(\pi) = x \in T_n$. Since

$$\pi_A = 20\ 18\ 13\ 11, \quad \pi_B = 9\ 7\ 4\ 2, \quad \text{and} \quad \pi_C = 5\ 6,$$

we have $x_i = A$ for $i \in \{1, 3, 8, 10\}$, $x_i = B$ for $i \in \{2, 4, 7, 9\}$, and $x_i = C$ for $i \in \{5, 6\}$. Thus

$$\Delta(\pi) = x = ABABCCBABA.$$

Consider $\Theta(\Delta(\pi)) \in \hat{P}_D(\emptyset; 10)$. Since $\Delta(\pi) = x = ABABCCBABA$, the lists \mathcal{A} , \mathcal{B} and \mathcal{C} are defined as

$$\mathcal{A} = [1 < 3 < 8 < 10], \quad \mathcal{B} = [9 > 7 > 4 > 2], \quad \text{and} \quad \mathcal{C} = [5 < 6].$$

Using the definition of Θ , we get

$$\Theta(\Delta(\pi)) = \Theta(x) = 20\ 18\ 13\ 11\ 9\ 7\ 4\ 2\ 5\ 6 \mid 15\ 16\ 19\ 17\ 14\ 12\ 10\ 8\ 3\ 1 = \pi.$$

5. Questions and future work

We end this paper with a few questions of interest. We suspect that the sets we call pattern bundles have appeared elsewhere in the literature on Coxeter groups, but we do not know of such a reference. (Note that the pattern bundles are the fibers of an order-preserving flattening map from \mathcal{C}_n to \mathfrak{S}_n that differs from the usual 2^n to 1

projection of signed permutations to \mathfrak{S}_n , which forgets the negative signs.) If these sets have not been studied before, then our first question is:

Problem 1. Can the pattern bundles of types C_n and D_n be used to study other permutation statistics (such as descent sets for instance)?

We can also ask whether these techniques can be applied to study other groups having *suitably nice* embeddings into \mathfrak{S}_N , and whether the peak set of the image encodes any information about the embedded group.

Problem 2. Can the methods used in this paper be applied to study peak sets of groups such as the dihedral groups or Coxeter groups of exceptional type by embedding them into \mathfrak{S}_N for some N ?

We provide recursive formulas for the quantities $|\hat{P}_C(S; n)|$ and $|\hat{P}_D(S; n)|$ in Theorem 26 that can be used to find closed formulas for any particular choice of peak set S . Several of the special cases we consider in this paper give closed formulas for integer sequences appearing in [OEIS 1996]. Hence we believe the following would be an interesting undergraduate student research project.

Problem 3 (undergraduate student research project). Can one compute closed formulas for some families of peak sets and analyze which of these appear on the OEIS?

This leads us to our final question:

Problem 4. Can one discover closed combinatorial formulas for $|\hat{P}_C(S; n)|$ and $|\hat{P}_D(S; n)|$ in general?

Acknowledgements

The authors would like to thank the *Underrepresented Students in Topology and Algebra Symposium* (USTARS); if not for our chance encounter at USTARS this collaboration may not have materialized! We also thank Sara Billey, Christophe Hohlweg, and Bruce Sagan for helpful conversations about this paper. Pamela E. Harris gratefully acknowledges travel support from the Photonics Research Center and the Mathematical Sciences Center of Excellence at the United States Military Academy.

References

- [Aguiar et al. 2004] M. Aguiar, N. Bergeron, and K. Nyman, “The peak algebra and the descent algebras of types B and D ”, *Trans. Amer. Math. Soc.* **356**:7 (2004), 2781–2824. MR Zbl
- [Aguiar et al. 2006a] M. Aguiar, N. Bergeron, and F. Sottile, “Combinatorial Hopf algebras and generalized Dehn–Sommerville relations”, *Compos. Math.* **142**:1 (2006), 1–30. MR Zbl
- [Aguiar et al. 2006b] M. Aguiar, K. Nyman, and R. Orellana, “New results on the peak algebra”, *J. Algebraic Combin.* **23**:2 (2006), 149–188. MR Zbl
- [Bergeron and Hohlweg 2006] N. Bergeron and C. Hohlweg, “Coloured peak algebras and Hopf algebras”, *J. Algebraic Combin.* **24**:3 (2006), 299–330. MR Zbl

- [Bergeron and Sottile 2002] N. Bergeron and F. Sottile, “Skew Schubert functions and the Pieri formula for flag manifolds”, *Trans. Amer. Math. Soc.* **354**:2 (2002), 651–673. MR Zbl
- [Bergeron et al. 2000] N. Bergeron, S. Mykytiuk, F. Sottile, and S. van Willigenburg, “Noncommutative Pieri operators on posets”, *J. Combin. Theory Ser. A* **91**:1–2 (2000), 84–110. MR Zbl
- [Bergeron et al. 2002] N. Bergeron, S. Mykytiuk, F. Sottile, and S. van Willigenburg, “Shifted quasi-symmetric functions and the Hopf algebra of peak functions”, *Discrete Math.* **246**:1–3 (2002), 57–66. MR Zbl
- [Billera et al. 2003] L. J. Billera, S. K. Hsiao, and S. van Willigenburg, “Peak quasisymmetric functions and Eulerian enumeration”, *Adv. Math.* **176**:2 (2003), 248–276. MR Zbl
- [Billley and Haiman 1995] S. Billey and M. Haiman, “Schubert polynomials for the classical groups”, *J. Amer. Math. Soc.* **8**:2 (1995), 443–482. MR Zbl
- [Billley and Lakshmibai 2000] S. Billey and V. Lakshmibai, *Singular loci of Schubert varieties*, Progress in Mathematics **182**, Birkhäuser, Boston, 2000. MR Zbl
- [Billley et al. 2013] S. Billey, K. Burdzy, and B. E. Sagan, “Permutations with given peak set”, *J. Integer Seq.* **16**:6 (2013), Article 13.6.1, 18. MR Zbl
- [Billley et al. 2015] S. Billey, K. Burdzy, S. Pal, and B. E. Sagan, “On meteors, earthworms and WIMPs”, *Ann. Appl. Probab.* **25**:4 (2015), 1729–1779. MR Zbl
- [Billley et al. 2016] S. Billey, M. Fahrback, and A. Talmage, “Coefficients and roots of peak polynomials”, *Exp. Math.* **25**:2 (2016), 165–175. MR Zbl
- [Björner and Brenti 2005] A. Björner and F. Brenti, *Combinatorics of Coxeter groups*, Graduate Texts in Mathematics **231**, Springer, New York, 2005. MR Zbl
- [Castro-Velez et al. 2013] F. Castro-Velez, A. Diaz-Lopez, R. Orellana, J. Pastrana, and R. Zevallos, “Number of permutations with same peak set for signed permutations”, preprint, 2013. To appear in *J. Comb.* arXiv
- [Diaz-Lopez et al. 2015] A. Diaz-Lopez, P. E. Harris, E. Insko, and D. Perez-Lavin, “Peaks Sets of Classical Coxeter Groups”, preprint, 2015. arXiv
- [Kasraoui 2012] A. Kasraoui, “The most frequent peak set of a random permutation”, preprint, 2012. arXiv
- [Nyman 2003] K. L. Nyman, “The peak algebra of the symmetric group”, *J. Algebraic Combin.* **17**:3 (2003), 309–322. MR Zbl
- [OEIS 1996] OEIS, “The on-line encyclopedia of integer sequences”, 1996, <http://oeis.org>.
- [Petersen 2007] T. K. Petersen, “Enriched P -partitions and peak algebras”, *Adv. Math.* **209**:2 (2007), 561–610. MR Zbl
- [Stanley 2012] R. P. Stanley, *Enumerative combinatorics, Volume 1*, 2nd ed., Cambridge Studies in Advanced Mathematics **49**, Cambridge University Press, 2012. MR Zbl
- [Stembridge 1997] J. R. Stembridge, “Enriched P -partitions”, *Trans. Amer. Math. Soc.* **349**:2 (1997), 763–788. MR Zbl

Received: 2015-09-11 Revised: 2016-01-21 Accepted: 2016-02-07

adiazlo1@swarthmore.edu

*Department of Mathematics and Statistics,
Swarthmore College, Swarthmore, PA 19081, United States*

peh2@williams.edu

*Department of Mathematics and Statistics, Williams College,
Williamstown, MA 01267, United States*

einsko@fgcu.edu

*Department of Mathematics, Florida Gulf Coast University,
Fort Myers, FL 33965, United States*

darleenpl@uky.edu

*Department of Mathematics, University of Kentucky,
Lexington, KY 40506, United States*

Fox coloring and the minimum number of colors

Mohamed Elhamdadi and Jeremy Kerr

(Communicated by Kenneth S. Berenhaut)

We study Fox colorings of knots that are 13-colorable. We prove that any 13-colorable knot has a diagram that uses exactly five of the thirteen colors that are assigned to the arcs of the diagram. Due to an existing lower bound, this gives that the minimum number of colors of any 13-colorable knot is 5.

1. Introduction

Fox [1962] introduced a diagrammatic definition of colorability of a knot K by \mathbb{Z}_m (the integers modulo m). This notion of colorability is clearly one of the simplest invariants of knots. For a natural number m greater than 1, a diagram D of a knot K is m -colorable if at every crossing, the sum of the colors of the under-arcs is twice the color of the over-arc (modulo m), as in Figure 1.

It is well known [Fox 1962] that for a prime p , a knot K is p -colorable if and only if p divides the determinant of K . The problem of finding the minimum number of colors for p -colorable knots with p prime and less than or equal to 11 was studied in [Satoh 2009; Oshiro 2010; Lopes and Matias 2012; Hayashi et al. 2012]. For example, Satoh [2009] proved that any 5-colorable knot admits a nontrivially 5-colored diagram where the coloring assignment uses only four of the five available colors. For a prime p , let K be a p -colorable knot and let $C_p(K)$ denote the minimum number of colors among all diagrams of the knot K . In [Nakamura et al. 2013], it was proved that $C_p(K) \geq \lfloor \log_2 p \rfloor + 2$. This implies that in our case, $p = 13$, the minimum number of colors of 13-colorable knots is greater than or equal to 5. In fact, the goal of this article is to prove equality, that is, $C_{13}(K) = 5$.

2. Fox coloring and the minimum number of colors of 13-colorable knots

Notation. We use $\{a|b|c\}$ to denote a crossing, as in Figure 1, where a and c are the colors of the under-arcs, b is the color of the over-arc and $a + c \equiv 2b \pmod{13}$. When the crossing is of the type $\{c|c|c\}$ (trivial coloring), we will omit over- and under-crossings and draw the arcs crossing each other.

MSC2010: 57M25.

Keywords: knots, fox colorings, minimum number of colors.

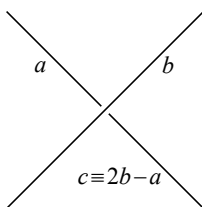


Figure 1

Theorem 2.1. *Any 13-colorable knot has a 13-colored diagram with exactly five colors. Thus, $C_{13}(K) = 5$ for any 13-colorable knot K .*

Proof. We prove this theorem using eight lemmas. In each of the following lemmas we decrease the coloring scheme of the diagram by one color c . To accomplish this we first transform any crossings of the form $\{c|c|c\}$. That is, when c is both an over-arc and an under-arc, we remove c as an over-arc by transforming any crossings of the form $\{a|c|2c-a\}$, where $a \in \mathbb{Z}_{13} \setminus \{c\}$. Finally, we complete each lemma by removing c as an under-arc in a case-by-case method. In these under-arc cases we must consider when c connects two crossings of the same color and when c connects two crossings of different colors. □

Eliminating the color 12.

Lemma 2.2. *Any 13-colorable knot has a 13-colored diagram D with no arc colored by 12.*

Proof. Let $c = 12$. We first transform any crossing of the form $\{12|12|12\}$. If there is any such crossing, there is an adjacent crossing of the form $\{12|a|2a+1\}$ or $\{a|12|11-a\}$, where $a \in \mathbb{Z}_{13} \setminus \{12\}$. In either case, since $11 - a \neq 12$ and $2a + 1 \neq 12$ for any a in $\mathbb{Z}_{13} \setminus \{12\}$, we transform the diagram as in Figures 2 and 3.

Next, we remove 12 as an over-arc by transforming any crossings of the form $\{a|12|11-a\}$. Since $2a + 1 \neq 12$ and $3a + 2 \neq 12$ for any $a \in \mathbb{Z}_{13} \setminus \{12\}$, we transform the diagram as in Figure 4.

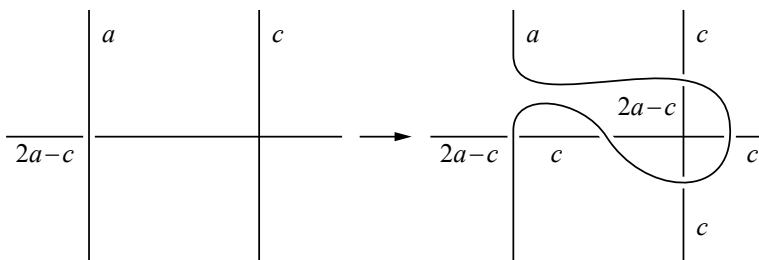


Figure 2

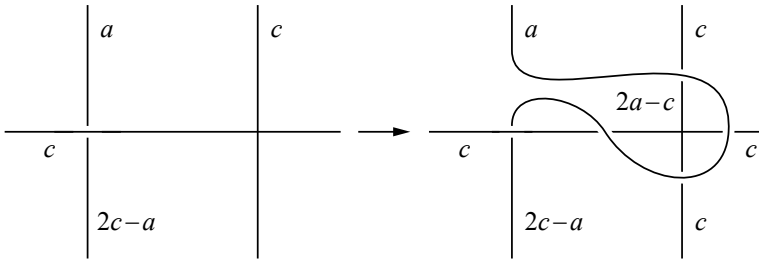


Figure 3

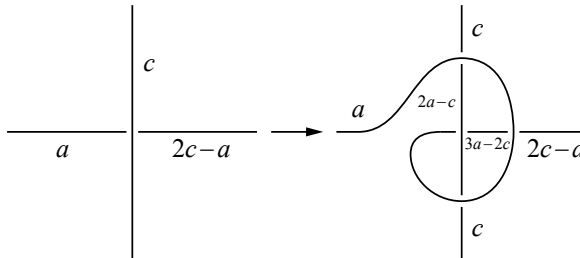


Figure 4

We complete the proof of the lemma by removing 12 as an under-arc in a case-by-case method. We first consider the case where 12 is an under-arc connecting two crossings of the form $\{12|a|2a+1\}$. Since $2a + 1 \neq 12$, $3a + 2 \neq 12$, and $4a + 3 \neq 12$ for any $a \in \mathbb{Z}_{13} \setminus \{12\}$, we transform the diagram as in Figure 5.

Now we consider the case where 12 is an under-arc connecting two crossings of the forms $\{2a+1|a|12\}$ and $\{12|2a+1|4a+3\}$. Since $2a + 1 \neq 12$ and $3a + 2 \neq 12$ for any $a \in \mathbb{Z}_{13} \setminus \{12\}$, we transform the diagram as in Figure 6.

Lastly we consider the case where 12 is an under-arc connecting two crossings of the forms $\{2a+1|a|12\}$ and $\{12|b|2b+1\}$, where $a \neq b$ and $b \neq 2a + 1$ for any a and b in $\mathbb{Z}_{13} \setminus \{12\}$. Since $2a - 2b - 1 \neq 12$ and $2a - b \neq 12$ for any a and b in $\mathbb{Z}_{13} \setminus \{12\}$ (from $a \neq b$ and $b \neq 2a + 1$ respectively), we transform the diagram as in Figure 7. \square

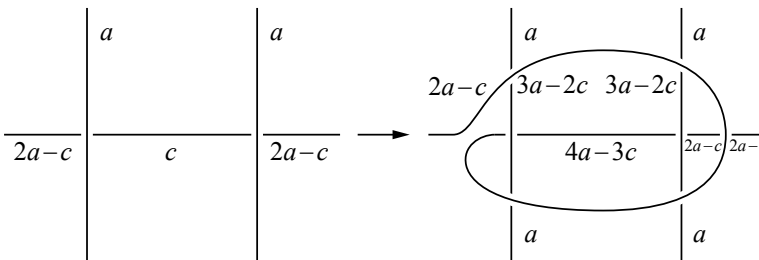


Figure 5

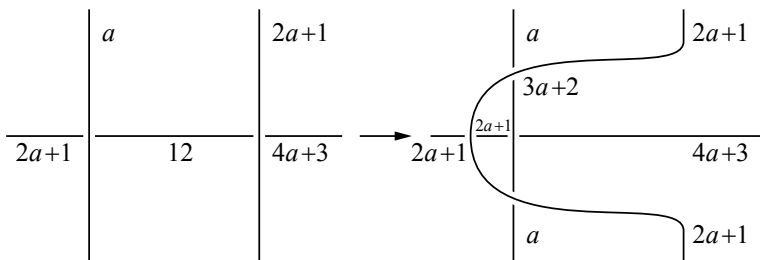


Figure 6

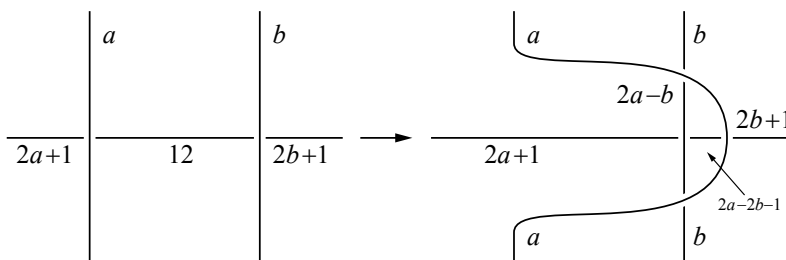


Figure 7

Eliminating the color 11.

Lemma 2.3. Any 13-colorable knot has a 13-colored diagram D with no arc colored by 11 or 12.

Proof. Let $c = 11$. By the previous lemma we assume that no arc in D is colored by 12. We first transform any crossing of the form $\{11|11|11\}$. If there is any such crossing, there is an adjacent crossing of the form $\{11|a|2a+2\}$ or $\{a|11|9-a\}$, where a is in $\mathbb{Z}_{13} \setminus \{11, 12\}$. If $a \neq 5, 10$, then $9 - a \neq 11, 12$ and $2a + 2 \neq 11, 12$ for any a in $\mathbb{Z}_{13} \setminus \{5, 10, 11, 12\}$, so we transform the diagram as in Figures 2 and 3.

If $a = 5$ as an under-arc, we transform the diagram as in Figure 8. Now, a cannot equal 5 as an over-arc, otherwise $2a + 2 = 12$, contradicting our assumption that no arc is colored by 12.

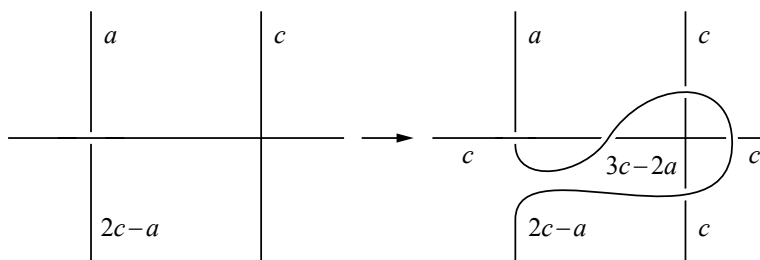


Figure 8

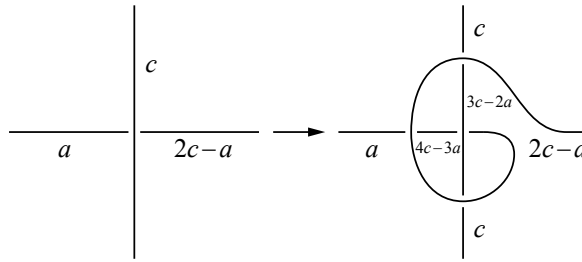


Figure 9

If $a = 10$ as an over-arc, we transform the diagram as in Figure 2. Similarly a cannot equal 10 as an under-arc, otherwise $9 - a = 12$, which is a contradiction.

Next, we remove 11 as an over-arc by transforming any crossings of the form $\{a | 11 | 9 - a\}$. Since $9 - a \neq 11, 12$, we have $a \neq 10$. Therefore if $a \neq 5, 7$ then $2a + 2 \neq 11, 12$ and $3a + 4 \neq 11, 12$ for any a in $\mathbb{Z}_{13} \setminus \{5, 7, 10, 11, 12\}$, and we transform the diagram as in Figure 4. If $a = 5$ or $a = 7$, we transform the diagram as in Figure 9.

We complete the proof of the lemma by removing 11 as an under-arc in a case-by-case method. We first consider the case where 11 is an under-arc connecting two crossings of the form $\{11 | a | 2a + 2\}$. Since $2a + 2 \neq 11, 12$, we have $a \neq 5$. If $a \neq 7, 8$, then $3a + 4 \neq 11, 12$ and $4a + 6 \neq 11, 12$ for any a in $\mathbb{Z}_{13} \setminus \{5, 7, 8, 11, 12\}$, and we transform the diagram as in Figure 5. If $a = 7$, we transform the diagram as in Figure 10. If $a = 8$, we transform the diagram as in Figure 11.

Now we consider the case where 11 is an under-arc connecting two crossings of the forms $\{2a + 2 | a | 11\}$ and $\{11 | b | 2b + 2\}$, where $a \neq b$ for any a and b in $\mathbb{Z}_{13} \setminus \{5, 11, 12\}$. (Note $a, b \neq 5$, otherwise $2a + 2 = 12$ or $2b + 2 = 12$.)

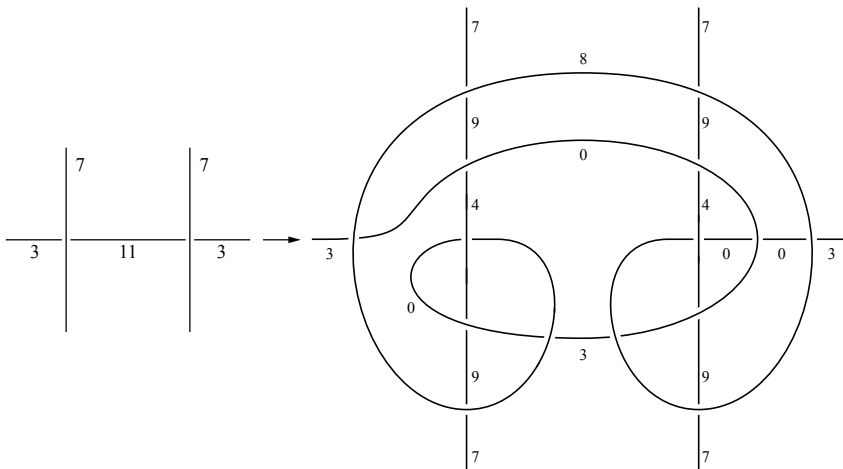


Figure 10

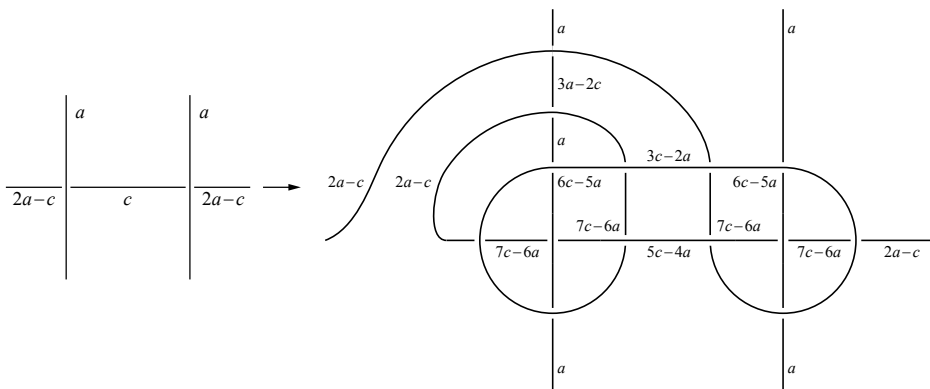


Figure 11

If $(a, b) \neq (0, 6), (6, 0), (3, 7), (7, 3)$ then either

$$2a - 2b - 2 \neq 11, 12 \quad \text{and} \quad 2a - b \neq 11, 12$$

or

$$2b - 2a - 2 \neq 11, 12 \quad \text{and} \quad 2b - a \neq 11, 12$$

for any a and b in $\mathbb{Z}_{13} \setminus \{5, 11, 12\}$, and we transform the diagram as in Figure 12.

If $(a, b) = (0, 6)$, we transform the diagram as in Figure 13. A similar transformation works for the case $(a, b) = (6, 0)$.

If $(a, b) = (3, 7)$, we transform the diagram as in Figure 14. A similar transformation works for the case $(a, b) = (7, 3)$. □

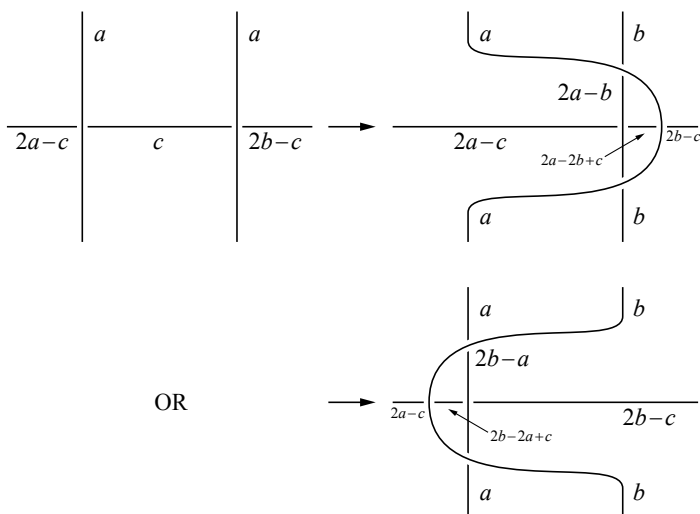


Figure 12

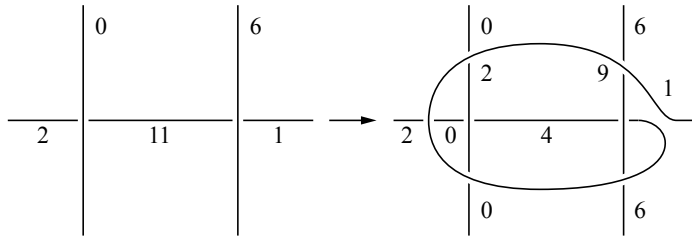


Figure 13

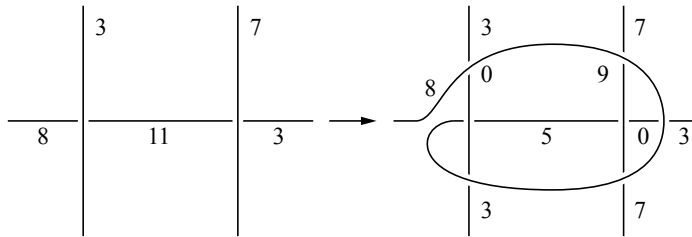


Figure 14

Eliminating the color 7.

Lemma 2.4. Any 13-colorable knot has a 13-colored diagram D with no arc colored by 7, 11, or 12.

Proof. Let $c = 7$. By the previous lemmas we assume that no arc in D is colored by 11 or 12. We first transform any crossing of the form $\{7|7|7\}$. If there is any such crossing, there is an adjacent crossing of the form $\{7|a|2a+6\}$ or $\{a|7|1-a\}$, where a is in $\mathbb{Z}_{13} \setminus \{7, 11, 12\}$. If $a \neq 2, 3, 9$ then $1-a \neq 7, 11, 12$ and $2a+6 \neq 7, 11, 12$ for any a in $\mathbb{Z}_{13} \setminus \{2, 3, 7, 9, 11, 12\}$, so we transform the diagram as in Figures 2 and 3.

If $a = 2$ as an over-arc, we transform the diagram as in Figure 2. Note a cannot equal 2 as an under-arc, otherwise $1-a = 12$, contradicting our assumption that no arc is colored by 12.

Now a cannot be 3 as an over-arc or an under-arc, otherwise $1-a = 11$ and $2a+6 = 12$, contradicting our assumption that no arc is colored by 11 or 12. If $a = 9$ as an under-arc, we transform the diagram as in Figure 8. Note a cannot equal 9 as an over-arc, otherwise $2a+6 = 11$, contradicting our assumption that no arc is colored by 11. Therefore any crossings of the form $\{7|7|7\}$ are removed.

Next, we remove 7 as an over-arc by transforming any crossings of the form $\{a|7|1-a\}$. Since $1-a \neq 7, 11, 12$, we have $a \neq 2, 3$. Therefore if $a \neq 0, 4, 9$ then $2a+6 \neq 7, 11, 12$ and $3a+12 \neq 7, 11, 12$ for any a in $\mathbb{Z}_{13} \setminus \{0, 2, 3, 4, 7, 9, 11, 12\}$, and we transform the diagram as in Figure 4. If $a = 0, 4, 9$, we transform the diagram as in Figure 9.

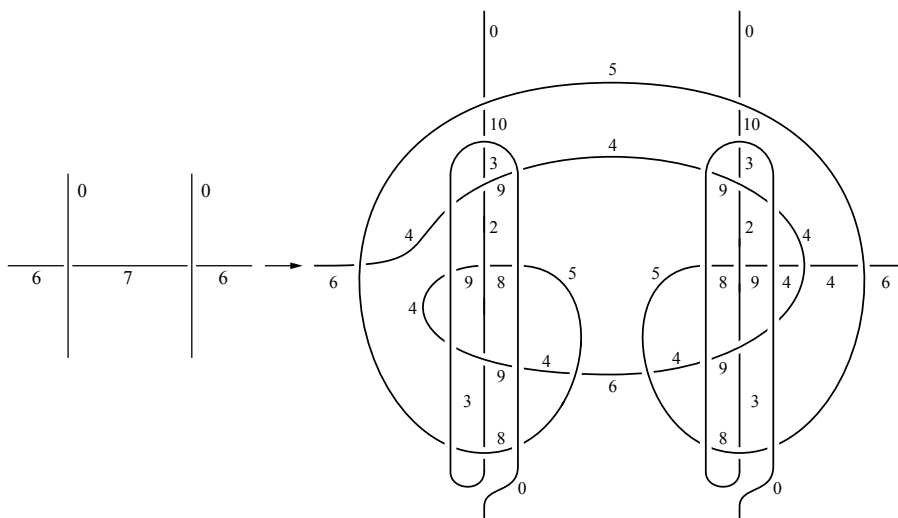


Figure 15

We complete the proof of the lemma by removing 7 as an under-arc in a case-by-case method. We first consider the case where 7 is an under-arc connecting two crossings of the form $\{7|a|2a+6\}$. Since $2a + 6 \neq 7, 11, 12$, we have $a \neq 3, 9$. If $a \neq 0, 4, 5, 8$ then $3a + 12 \neq 7, 11, 12$ and $4a + 5 \neq 7, 11, 12$ for any a in $\mathbb{Z}_{13} \setminus \{0, 3, 4, 5, 7, 8, 9, 11, 12\}$, and we transform the diagram as in Figure 5. If $a = 0$, we transform the diagram as in Figure 15. If $a = 4$, we transform the diagram as in Figure 16. If $a = 5$, we transform the diagram as in Figure 17. If $a = 8$, we transform the diagram as in Figure 11.

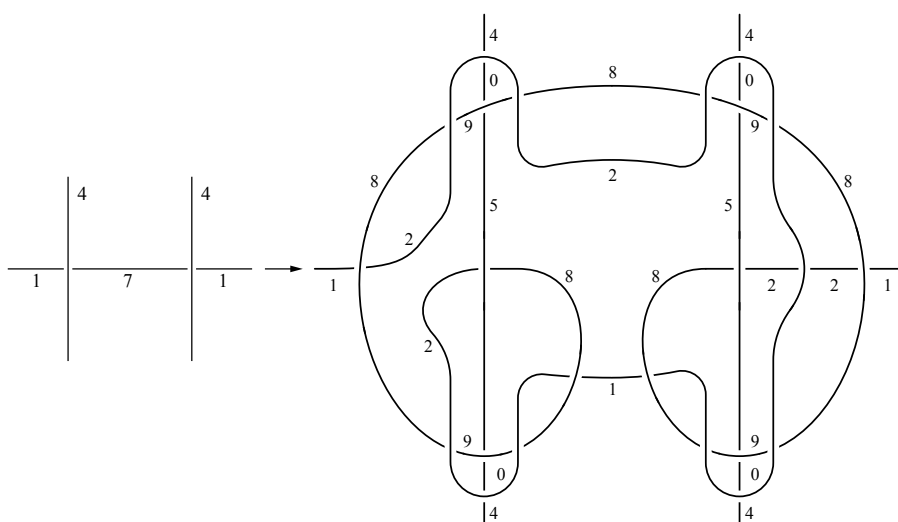


Figure 16

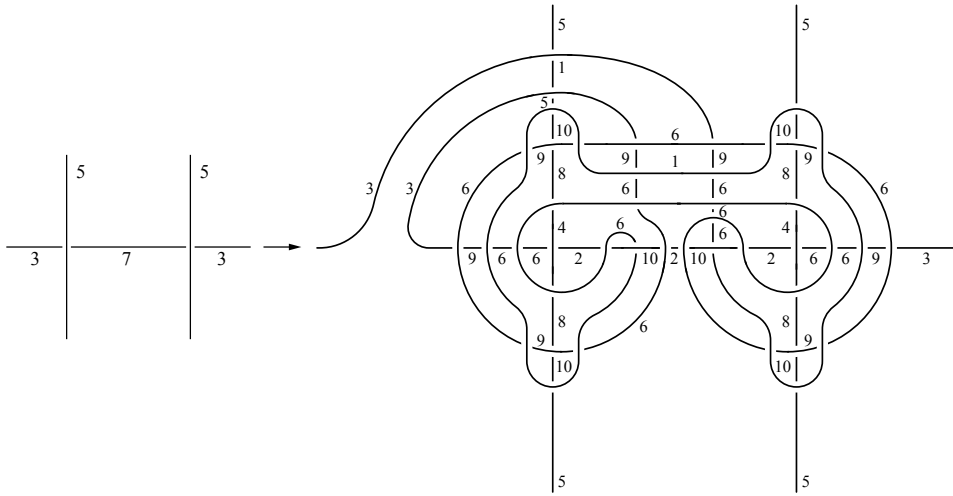


Figure 17

Now we consider the case where 7 is an under-arc connecting two crossings of the forms $\{2a+6|a|7\}$ and $\{7|b|2b+6\}$, where $a \neq b$ for any a and b in $\mathbb{Z}_{13} \setminus \{3, 7, 9, 11, 12\}$. (Note $a, b \neq 3, 9$, otherwise $2a + 6 = 11, 12$ or $2b + 6 = 11, 12$.) If

$$(a, b) \neq (0, 2), (2, 0), (0, 6), (6, 0), (1, 4), (4, 1), (4, 8), (8, 4)$$

then either

$$2a - 2b - 6 \neq 7, 11, 12 \quad \text{and} \quad 2a - b \neq 7, 11, 12$$

or

$$2b - 2a - 6 \neq 7, 11, 12 \quad \text{and} \quad 2b - a \neq 7, 11, 12$$

for any a and b in $\mathbb{Z}_{13} \setminus \{3, 7, 9, 11, 12\}$, and we transform the diagram as in Figure 12.

If $(a, b) = (0, 2)$, we transform the diagram as in Figure 18. The case $(a, b) = (2, 0)$ is similar.

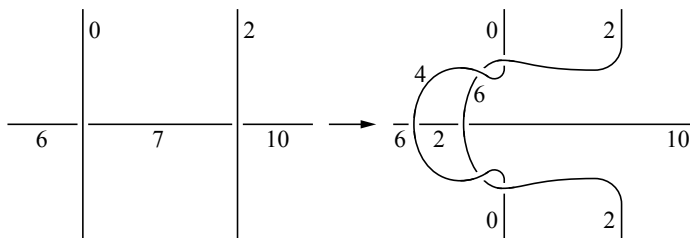


Figure 18

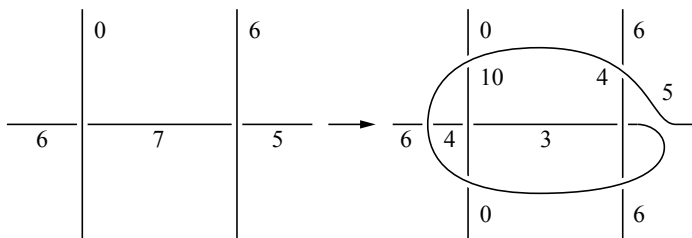


Figure 19

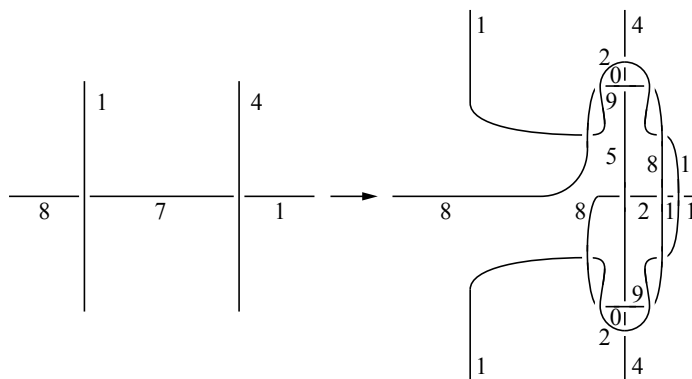


Figure 20

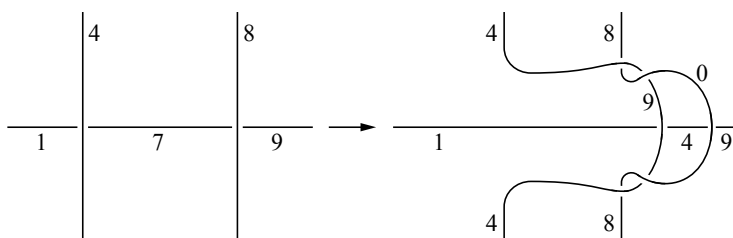


Figure 21

If $(a, b) = (0, 6)$, we transform the diagram as in Figure 19. The case $(a, b) = (6, 0)$ is similar.

If $(a, b) = (1, 4)$, we transform the diagram as in Figure 20. The case $(a, b) = (4, 1)$ is similar.

If $(a, b) = (4, 8)$, we transform the diagram as in following Figure 21. The case $(a, b) = (8, 4)$ is similar. □

Eliminating the color 8.

Lemma 2.5. Any 13-colorable knot has a 13-colored diagram D with no arc colored by 7, 8, 11, or 12.

Proof. Let $c = 8$. By the previous lemmas we assume that no arc in D is colored by 7, 11, or 12. We first transform any crossing of the form $\{8|8|8\}$. If there is any such crossing, there is an adjacent crossing of the form $\{8|a|2a+5\}$ or $\{a|8|3-a\}$, where a is in $\mathbb{Z}_{13} \setminus \{7, 8, 11, 12\}$. If $a \neq 1, 3, 4, 5, 9, 10$ then $3 - a \neq 7, 8, 11, 12$ and $2a + 5 \neq 7, 8, 11, 12$ for any a in $\mathbb{Z}_{13} \setminus \{1, 3, 4, 5, 7, 8, 9, 10, 11, 12\}$, so we transform the diagram as in Figures 2 and 3.

If $a = 4, 5, 9$ as an over-arc, we transform the diagram as in Figure 2. Note a cannot be 4, 5, or 9 as an under-arc, otherwise $3 - a = 7, 11, 12$, contradicting our assumption that no arc is colored by 7, 11, or 12. If $a = 1$ as an under-arc we transform the diagram as in Figure 8. Note a cannot be 1 as an over-arc, otherwise $2a + 5 = 7$, contradicting our assumption that no arc is colored by 7. If $a = 3$ as an under-arc, we transform the diagram as in Figure 8. Note a cannot be 3 as an over-arc, otherwise $2a + 5 = 11$, contradicting our assumption that no arc is colored by 11. If $a = 10$ as an under-arc, we transform the diagram as in Figure 8. Note a cannot be 10 as an over-arc, otherwise $2a + 5 = 12$, contradicting our assumption that no arc is colored by 12. Therefore any crossings of the form $\{8|8|8\}$ are removed.

Next, we remove 8 as an over-arc by transforming any crossings of the form $\{a|8|3-a\}$. Since $3 - a \neq 7, 8, 11, 12$, we have $a \neq 4, 5, 9$. Therefore if $a \neq 1, 3, 10$ then $2a + 5 \neq 7, 8, 11, 12$ and $3a + 10 \neq 7, 8, 11, 12$ for any a in $\mathbb{Z}_{13} \setminus \{1, 3, 4, 5, 7, 8, 9, 10, 11, 12\}$, and we transform the diagram as in Figure 4. If $a = 1, 3$ or 10 , we transform the diagram as in Figure 9.

We complete the proof of the lemma by removing 8 as an under-arc in a case-by-case method. We first consider the case where 8 is an under-arc connecting two crossings of the form $\{8|a|2a+5\}$. Since $2a + 5 \neq 7, 8, 11, 12$, we have $a \neq 1, 3, 10$. If $a \neq 5, 9$ then $3a + 10 \neq 7, 8, 11, 12$ and $4a + 2 \neq 7, 8, 11, 12$ for any a in $\mathbb{Z}_{13} \setminus \{1, 3, 5, 7, 8, 9, 10, 11, 12\}$, and we transform the diagram as in Figure 5. If $a = 5$, we transform the diagram as in Figure 22. If $a = 9$, we transform the diagram as in Figure 23.

Now we consider the case where 8 is an under-arc connecting two crossings of the forms $\{2a+5|a|8\}$ and $\{8|b|2b+5\}$, where $a \neq b$ for any a and b in $\mathbb{Z}_{13} \setminus \{1, 3, 7, 8, 10, 11, 12\}$. (Note $a, b \neq 1, 3, 10$, otherwise $2a + 5 = 7, 11, 12$ or $2b + 5 = 7, 11, 12$.) If $(a, b) \neq (0, 2), (2, 0), (0, 6), (6, 0), (2, 5), (5, 2)$ then either

$$2a - 2b - 5 \neq 7, 8, 11, 12 \quad \text{and} \quad 2a - b \neq 7, 8, 11, 12$$

or

$$2b - 2a - 5 \neq 7, 8, 11, 12 \quad \text{and} \quad 2b - a \neq 7, 8, 11, 12$$

for any a and b in $\mathbb{Z}_{13} \setminus \{1, 3, 7, 8, 10, 11, 12\}$, and we transform the diagram as in Figure 12.

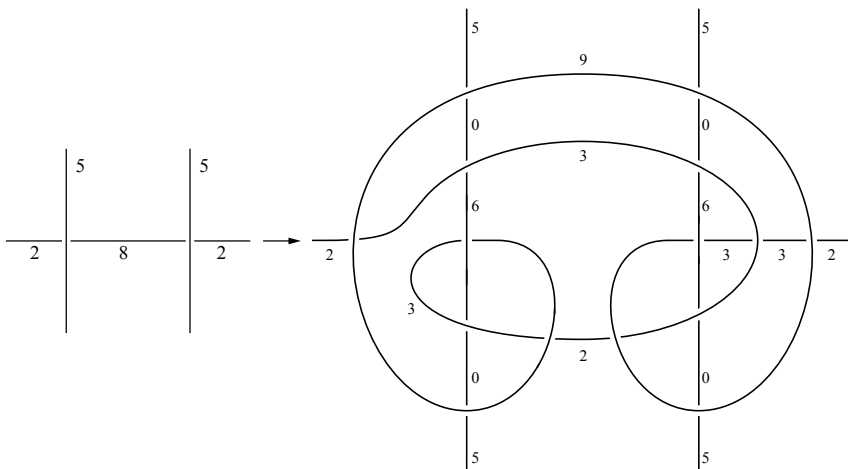


Figure 22

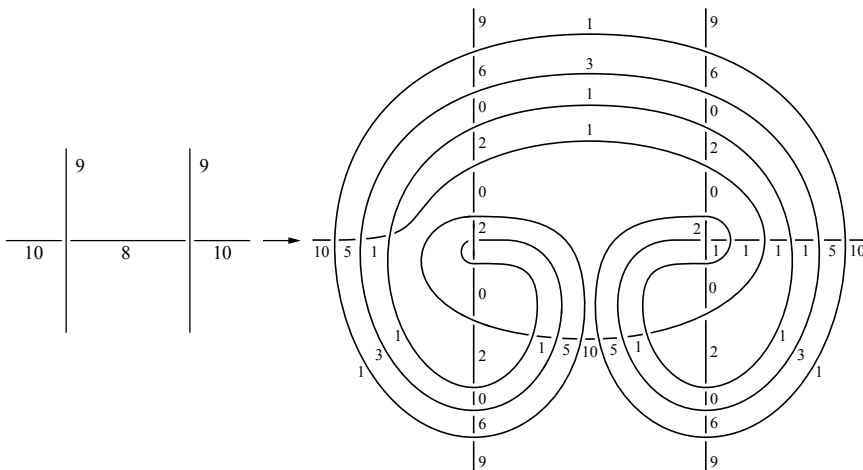


Figure 23

If $(a, b) = (0, 2)$, we transform the diagram as in Figure 24. The case $(a, b) = (2, 0)$ is similar.

If $(a, b) = (0, 6)$, we transform the diagram as in Figure 25. The case $(a, b) = (6, 0)$ is similar.

If $(a, b) = (2, 5)$, we transform the diagram as in Figure 26. The case $(a, b) = (5, 2)$ is similar. \square

Eliminating the color 6.

Lemma 2.6. Any 13-colorable knot has a 13-colored diagram D with no arc colored by 6, 7, 8, 11, or 12.

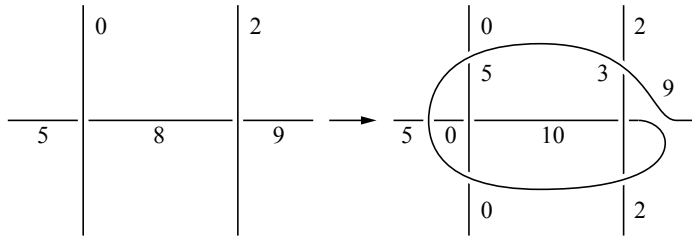


Figure 24

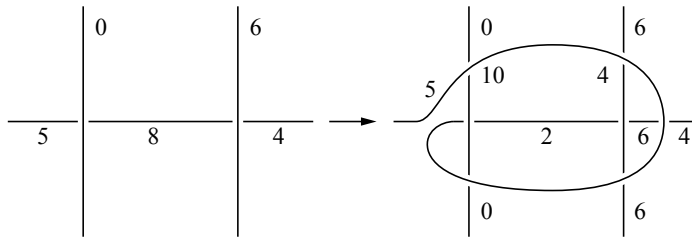


Figure 25

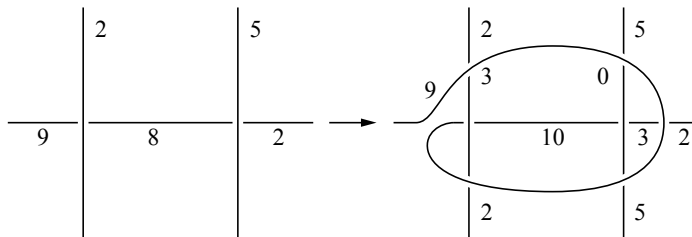


Figure 26

Proof. Let $c = 6$. By the previous lemmas we assume that no arc in D is colored by 7, 8, 11, or 12. We first transform any crossing of the form $\{6|6|6\}$. If there is any such crossing, there is an adjacent crossing of the form $\{6|a|2a+7\}$ or $\{a|6|12-a\}$, where a is in $\mathbb{Z}_{13} \setminus \{6, 7, 8, 11, 12\}$. With the exceptions of $a = 0, 2, 9$ as an over-arc (when $2a + 7 = 7, 8, 11, 12$) and $a = 0, 1, 4, 5$ as an under-arc (when $12 - a = 7, 8, 11, 12$), we transform the diagram as in Figures 2 and 3.

Now we must check when $a = 0, 2, 9$ as an under-arc. First and foremost a cannot equal 0 as an under-arc, otherwise $12 - a = 12$, contradicting our assumption that no arc is colored by 12. If $a = 2, 9$ as an under-arc, we transform the diagram as in Figure 8. Therefore any crossings of the form $\{6|6|6\}$ are removed.

Next, we remove 6 as an over-arc by transforming any crossings of the form $\{a|6|12-a\}$. Since $12 - a \neq 6, 7, 8, 11, 12$, we have $a \neq 0, 1, 4, 5$. With the exceptions of $a = 2, 9$ (when $2a + 7 = 6, 7, 8, 11, 12$ and $3a + 1 = 6, 7, 8, 11, 12$),

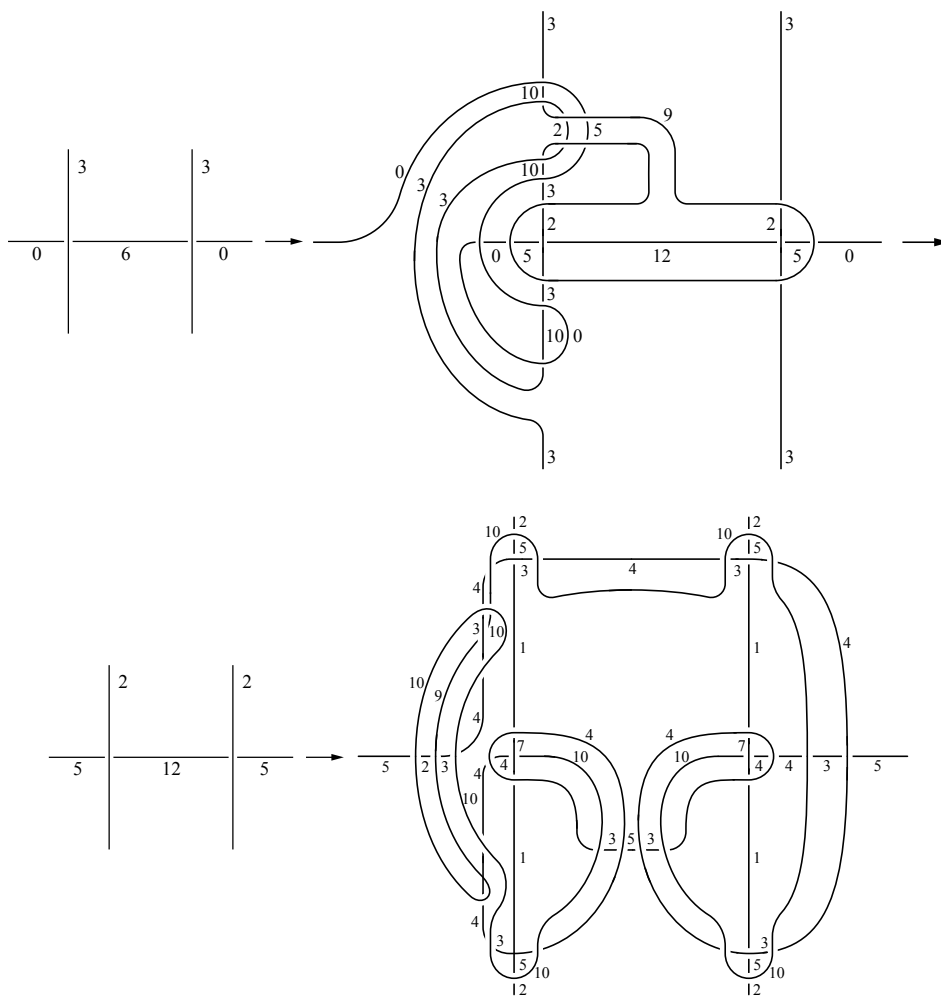


Figure 27

we transform the diagram as in Figure 4. If $a = 2$ or $a = 9$, we transform the diagram as in Figure 9.

We complete the proof of the lemma by removing 6 as an under-arc in a case-by-case method. We first consider the case where 6 is an under-arc connecting two crossings of the form $\{6|a|2a+7\}$. Since $2a + 7 \neq 6, 7, 8, 11, 12$, we have $a \neq 0, 2, 9$. If $a \neq 1, 3, 4$ then $3a + 1 \neq 6, 7, 8, 11, 12$ and $4a + 8 \neq 6, 7, 8, 11, 12$, so we transform the diagram as in Figure 5. If $a = 1$, we transform the diagram as in Figure 11. If $a = 3$, we transform the diagram as in Figures 27 and 28. If $a = 4$, we transform the diagram as in Figure 11.

Now we consider the case where 6 is an under-arc connecting two crossings of the forms $\{2a+7|a|6\}$ and $\{6|b|2b+7\}$, where $a \neq b$ for any a and b in

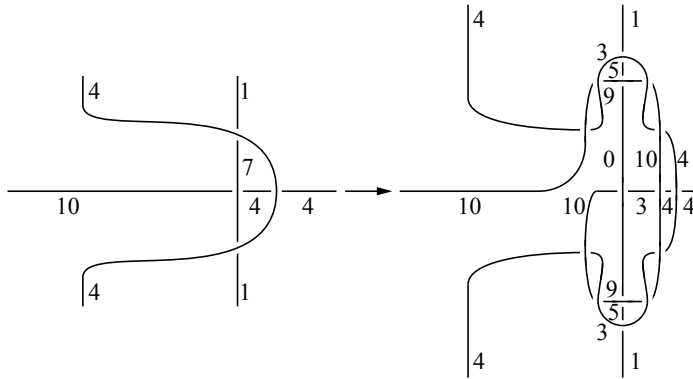


Figure 28

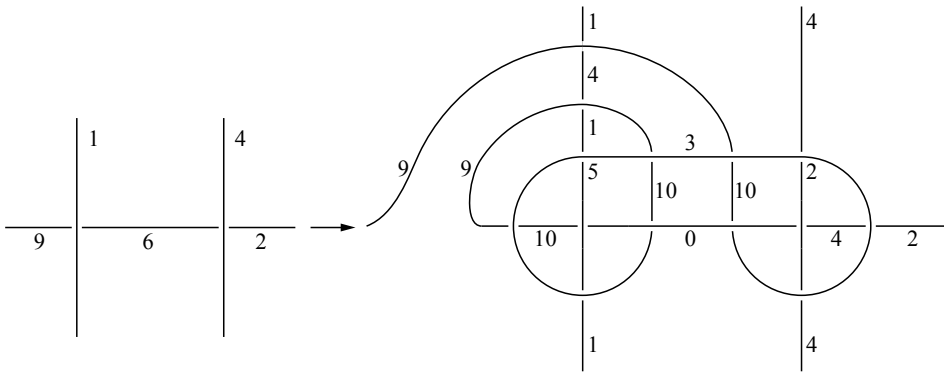


Figure 29

$\mathbb{Z}_{13} \setminus \{0, 2, 6, 7, 8, 9, 11, 12\}$. (Note $a, b \neq 0, 2, 9$, otherwise $2a + 7 = 7, 8, 11, 12$ or $2b + 7 = 7, 8, 11, 12$.)

If $(a, b) \neq (1, 4), (4, 1)$ then either

$$2a - 2b - 7 \neq 6, 7, 8, 11, 12 \quad \text{and} \quad 2a - b \neq 6, 7, 8, 11, 12$$

or

$$2b - 2a - 7 \neq 6, 7, 8, 11, 12 \quad \text{and} \quad 2b - a \neq 6, 7, 8, 11, 12$$

for any a and b in $\mathbb{Z}_{13} \setminus \{0, 2, 6, 7, 8, 9, 11, 12\}$, and we transform the diagram as in Figure 12.

If $(a, b) = (1, 4)$, we transform the diagram as in Figure 29. The case $(a, b) = (4, 1)$ is similar. \square

Eliminating the color 1.

Lemma 2.7. Any 13-colorable knot has a 13-colored diagram D with no arc colored by 1, 6, 7, 8, 11, or 12.

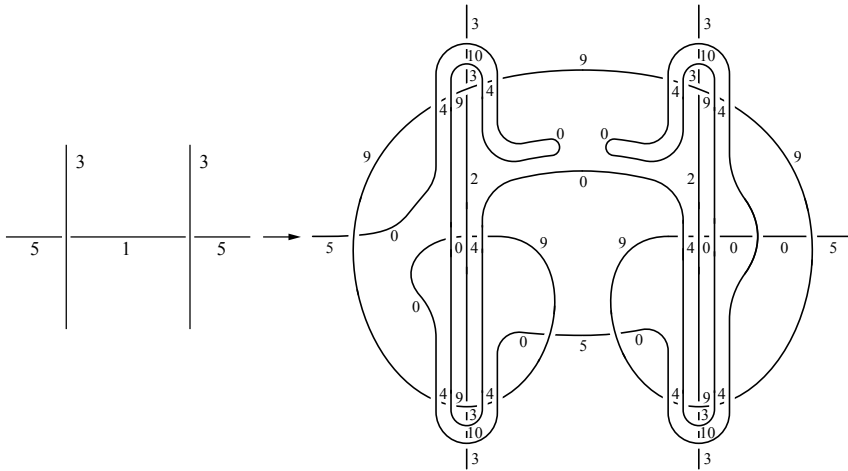


Figure 30

Proof. Let $c = 1$. By the previous lemmas we assume that no arc in D is colored by 6, 7, 8, 11, or 12. We first transform any crossing of the form $\{1|1|1\}$. If there is any such crossing, there is an adjacent crossing of the form $\{1|a|2a+12\}$ or $\{a|1|2-a\}$, where a is in $\mathbb{Z}_{13} \setminus \{1, 6, 7, 8, 11, 12\}$. With the exceptions of $a = 0, 4, 10$ as an over-arc (when $2a + 12 = 6, 7, 8, 11, 12$) and $a = 3, 4, 9$ as an under-arc (when $2 - a = 6, 7, 8, 11, 12$), we transform the diagram as in Figures 2 and 3.

Now we must check when $a = 0, 4, 10$ as an under-arc. We know a cannot be 4 as an under-arc, otherwise $2 - a = 11$, contradicting our assumption that no arc is colored by 11. If $a = 0$ or $a = 10$ as an under-arc, we transform the diagram as in Figure 8. Therefore any crossings of the form $\{1|1|1\}$ are removed.

Next, we remove 1 as an over-arc by transforming any crossings of the form $\{a|1|2-a\}$. Since $2-a \neq 1, 6, 7, 8, 11, 12$, we have $a \neq 3, 4, 9$. With the exceptions of $a = 0, 10$ (when $2a + 12 = 1, 6, 7, 8, 11, 12$ and $3a + 11 = 1, 6, 7, 8, 11, 12$), we transform the diagram as in Figure 4. If $a = 0$ or $a = 10$, we transform the diagram as in Figure 9.

We complete the proof by removing 1 as an under-arc in a case-by-case method. We first consider the case where 1 is an under-arc connecting two crossings of the form $\{1|a|2a+12\}$. Since $2a + 12 \neq 1, 6, 7, 8, 11, 12$, we have $a \neq 0, 4, 10$. If $a \neq 3, 9$ then $3a + 11 \neq 1, 6, 7, 8, 11, 12$ and $4a + 10 \neq 1, 6, 7, 8, 11, 12$, so we transform the diagram as in Figure 5. If $a = 3$, we transform the diagram as in Figure 30. If $a = 9$, we transform the diagram as in Figure 31.

Now we consider the case where 1 is an under-arc connecting two crossings of the forms $\{2a+12|a|1\}$ and $\{1|b|2b+12\}$, where $a \neq b$ for any a and b in $\mathbb{Z}_{13} \setminus \{0, 1, 4, 6, 7, 8, 10, 11, 12\}$. (Note $a, b \neq 0, 4, 10$, otherwise $2a + 12 = 1, 6, 7, 8, 11, 12$ or $2b + 12 = 1, 6, 7, 8, 11, 12$.) If $(a, b) \neq (2, 5), (5, 2), (3, 5), (5, 3)$ then

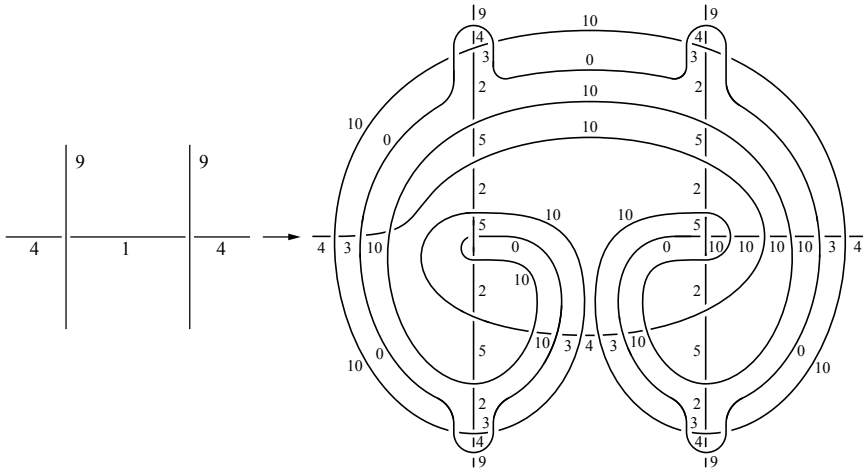


Figure 31

either

$$2a - 2b - 12 \neq 1, 6, 7, 8, 11, 12 \quad \text{and} \quad 2a - b \neq 1, 6, 7, 8, 11, 12$$

or

$$2b - 2a - 12 \neq 1, 6, 7, 8, 11, 12 \quad \text{and} \quad 2b - a \neq 1, 6, 7, 8, 11, 12$$

for any a and b in $\mathbb{Z}_{13} \setminus \{0, 1, 4, 6, 7, 8, 10, 11, 12\}$, and we transform the diagram as in Figure 12.

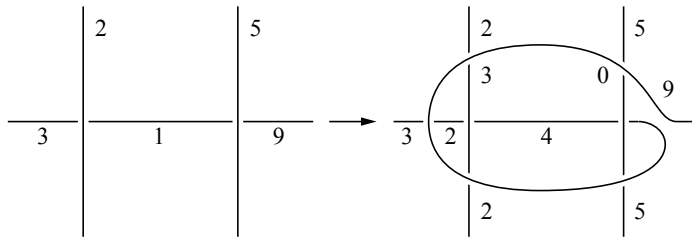


Figure 32

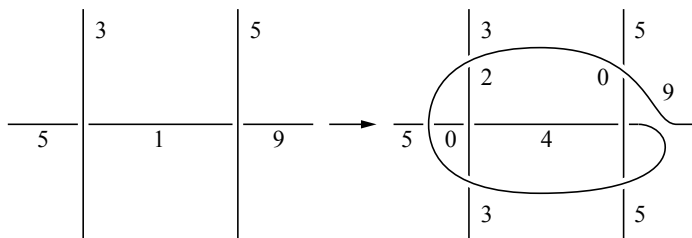


Figure 33

If $(a, b) = (2, 5)$, we transform the diagram as in Figure 32. The case $(a, b) = (5, 2)$ is similar.

If $(a, b) = (3, 5)$, we transform the diagram as in Figure 33. The case $(a, b) = (5, 3)$ is similar. \square

Eliminating the color 10.

Lemma 2.8. Any 13-colorable knot has a 13-colored diagram D with no arc colored by 1, 6, 7, 8, 10, 11, or 12.

Proof. Let $c = 10$. By the previous lemmas we assume that no arc in D is colored by 1, 6, 7, 8, 11, or 12. We first transform any crossing of the form $\{10|10|10\}$. If there is any such crossing, there is an adjacent crossing of the form $\{10|a|2a+3\}$ or $\{a|10|7-a\}$, where a is in $\mathbb{Z}_{13} \setminus \{1, 6, 7, 8, 10, 11, 12\}$. With the exceptions of $a = 2, 4, 9$ as an over-arc (when $2a + 3 = 1, 6, 7, 8, 11, 12$) and $a = 0, 9$ as an under-arc (when $7 - a = 1, 6, 7, 8, 11, 12$), we transform the diagram as in Figures 2 and 3.

Now we must check when $a = 2, 4, 9$ as an under-arc. We know a cannot be 9 as an under-arc, otherwise $7 - a = 11$, contradicting our assumption that no arc is colored by 11. If $a = 2$ or $a = 4$ as an under-arc, we transform the diagram as in Figure 8. Therefore any crossings of the form $\{10|10|10\}$ are removed.

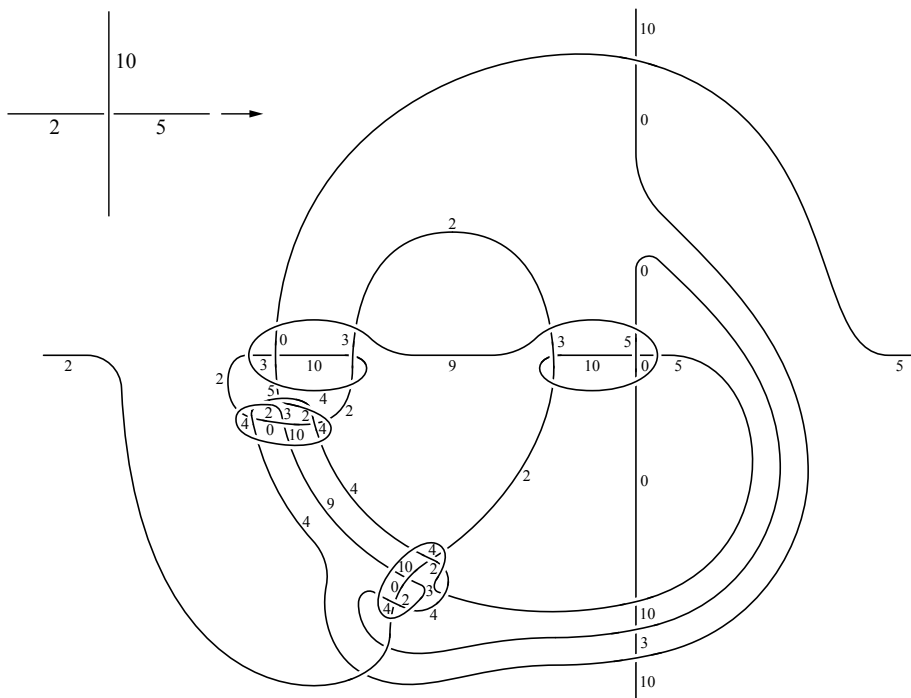


Figure 34

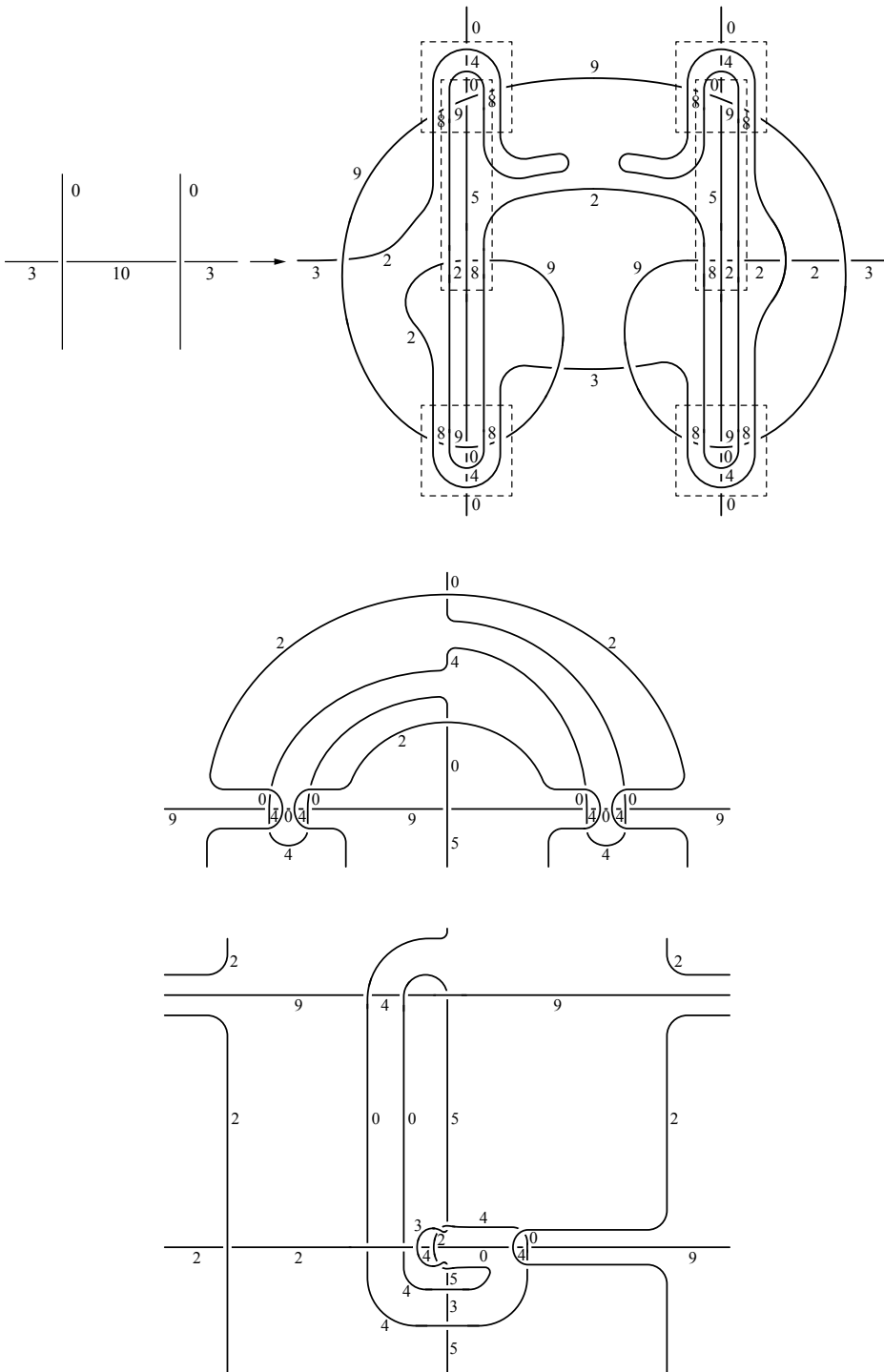


Figure 35

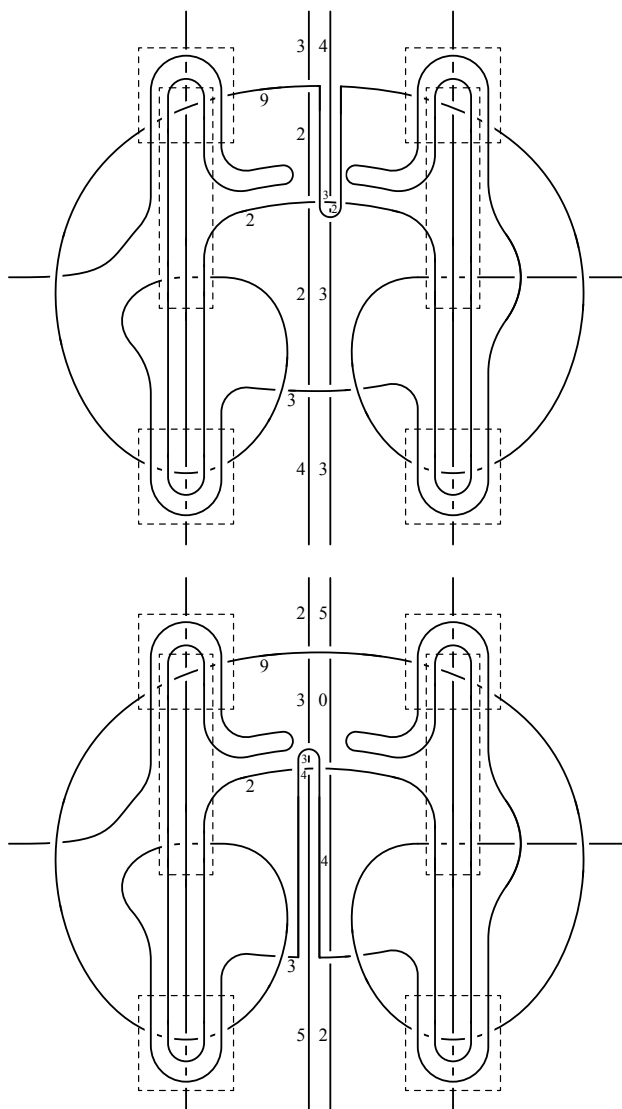


Figure 36

Next, we remove 10 as an over-arc by transforming any crossings of the form $\{a|10|7-a\}$. Since $7 - a \neq 1, 6, 7, 8, 10, 11, 12$, we have $a \neq 0, 9$. With the exceptions of $a = 2, 4, 5$ (when $2a + 3 = 1, 6, 7, 8, 10, 11, 12$ and $3a + 6 = 1, 6, 7, 8, 10, 11, 12$), we transform the diagram as in Figure 4. If $a = 2$, we transform the diagram as in Figure 34. If $a = 4$, we transform the diagram as in Figure 9. If $a = 5$, since $7 - a = 2$, we transform the diagram similarly to Figure 34, i.e., $a = 2$.

We complete the proof by removing 10 as an under-arc in a case-by-case method. We first consider the case where 10 is an under-arc connecting two crossings of the

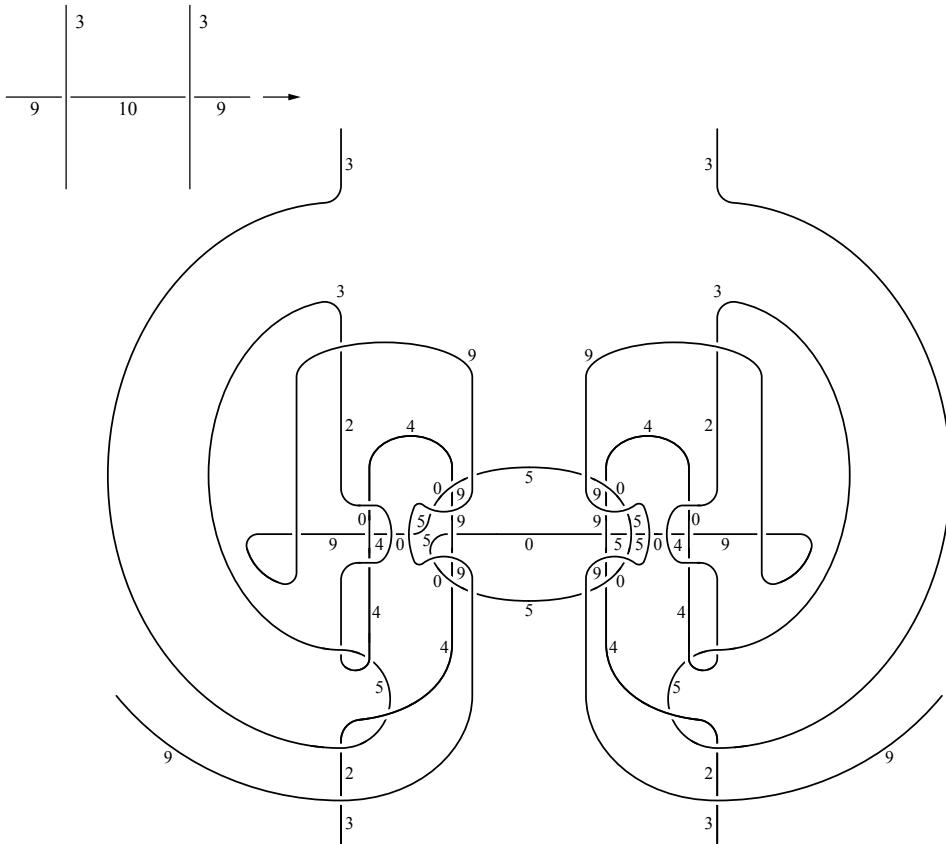


Figure 37

form $\{10|a|2a+3\}$. Since $2a + 3 \neq 1, 6, 7, 8, 10, 11, 12$, we have $a \neq 2, 4, 9$. So, we need to check $a = 0, 3, 5$. If $a = 0$, we transform the diagram as in Figure 35, and we shall refer to this transformation throughout Lemma 2.8. As such, two variations of this transformation are given in Figure 36. If $a = 3$, we transform the diagram as in Figure 37. If $a = 5$, we transform the diagram as in Figure 38. Note the center of $a = 5$ as well as the six dashed boxes are the same transformations we used for $a = 0$ and its variations. Also, there are two arcs colored by 10, each of which are transformed by $a = 3$ as in Figure 37.

Now we consider the case where 10 is an under-arc. There are six such cases: $(a, b) = (0, 3), (3, 0), (0, 5), (5, 0), (3, 5), (5, 3)$. If $(a, b) = (0, 3)$, we transform the diagram as in Figure 39. For eliminating the 10 arc, see the variations of $a = 0$ in Figure 36. The case $(a, b) = (3, 0)$ is similar.

If $(a, b) = (0, 5)$, we transform the diagram as in Figure 40. For eliminating the 10 arc, see $a = 5$ in Figure 38; however, we will use the variations of $a = 0$ in Figure 36 for the center. The case $(a, b) = (5, 0)$ is similar.

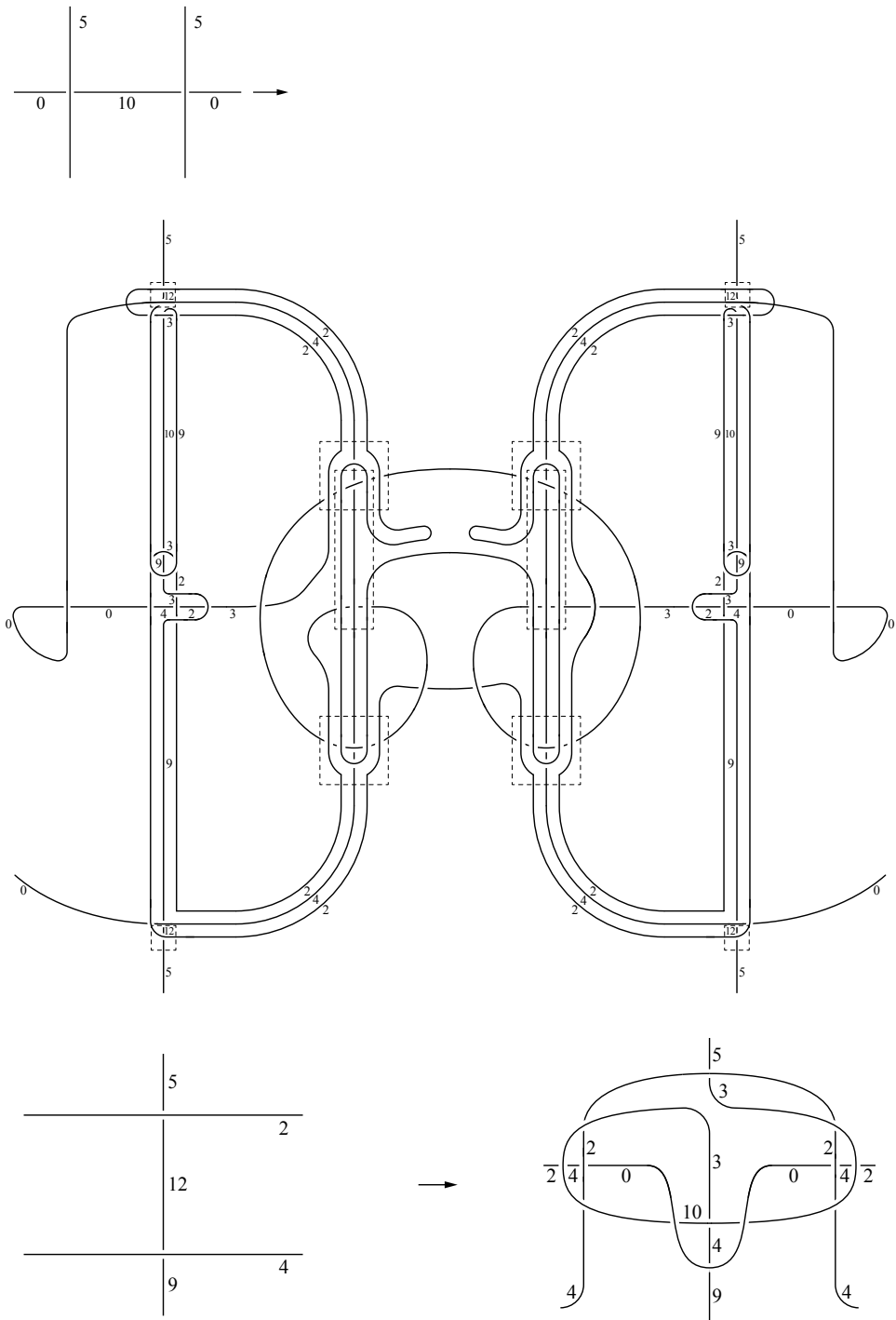


Figure 38

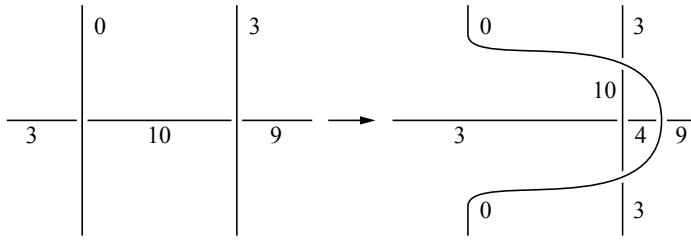


Figure 39

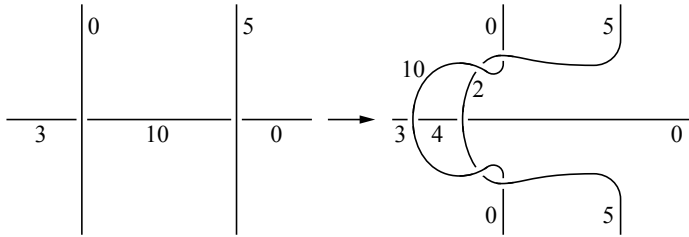


Figure 40

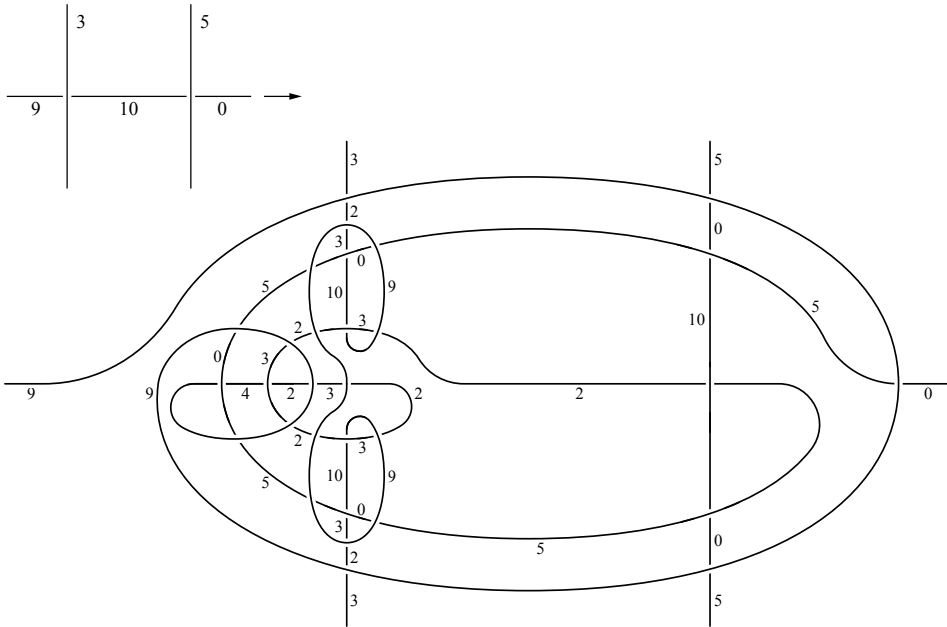


Figure 41

If $(a, b) = (3, 5)$, we transform the diagram as in Figure 41. For eliminating the 10 arcs, see the $(a, b) = (0, 3)$ case in Figure 39 and the $a = 5$ case in Figure 38 using the variations in Figure 36. The case $(a, b) = (5, 3)$ is similar. \square

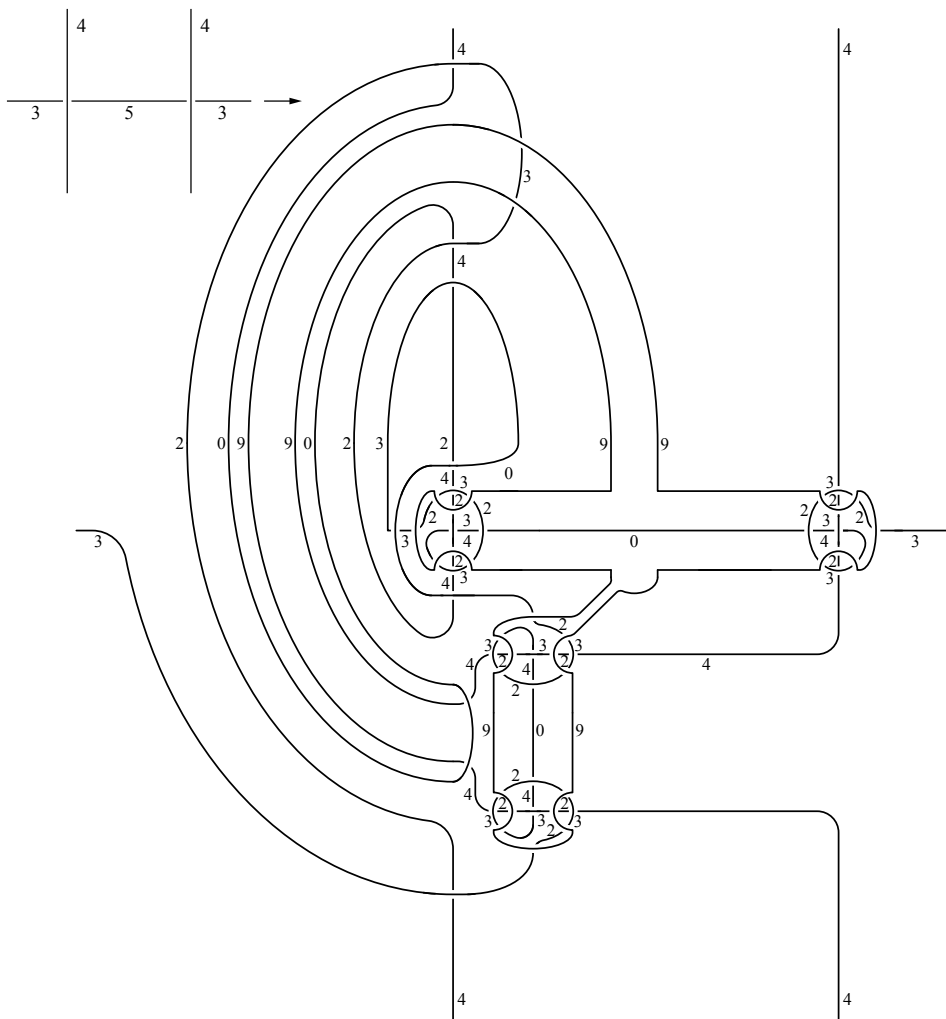


Figure 42

Eliminating the color 5.

Lemma 2.9. Any 13-colorable knot has a 13-colored diagram D with no arc colored by 1, 5, 6, 7, 8, 10, 11, or 12.

Proof. Let $c = 5$. By the previous lemmas we assume that no arc in D is colored by 1, 6, 7, 8, 10, 11, or 12. We first transform any crossing of the form $\{5|5|5\}$. If there is any such crossing, there is an adjacent crossing of the form $\{5|a|2a+8\}$, where a is in $\mathbb{Z}_{13} \setminus \{1, 5, 6, 7, 8, 10, 11, 12\}$. Since $10 - a = 1, 6, 7, 8, 10, 11, 12$ when $a = 0, 2, 3, 4, 9$, we know a cannot be an under-arc. Therefore, with the exceptions of $a = 0, 2, 3$ as an over-arc (when $2a + 8 = 1, 5, 6, 7, 8, 10, 11, 12$), we transform the diagram as in Figure 2. Therefore any crossings of the form $\{5|5|5\}$ are removed.

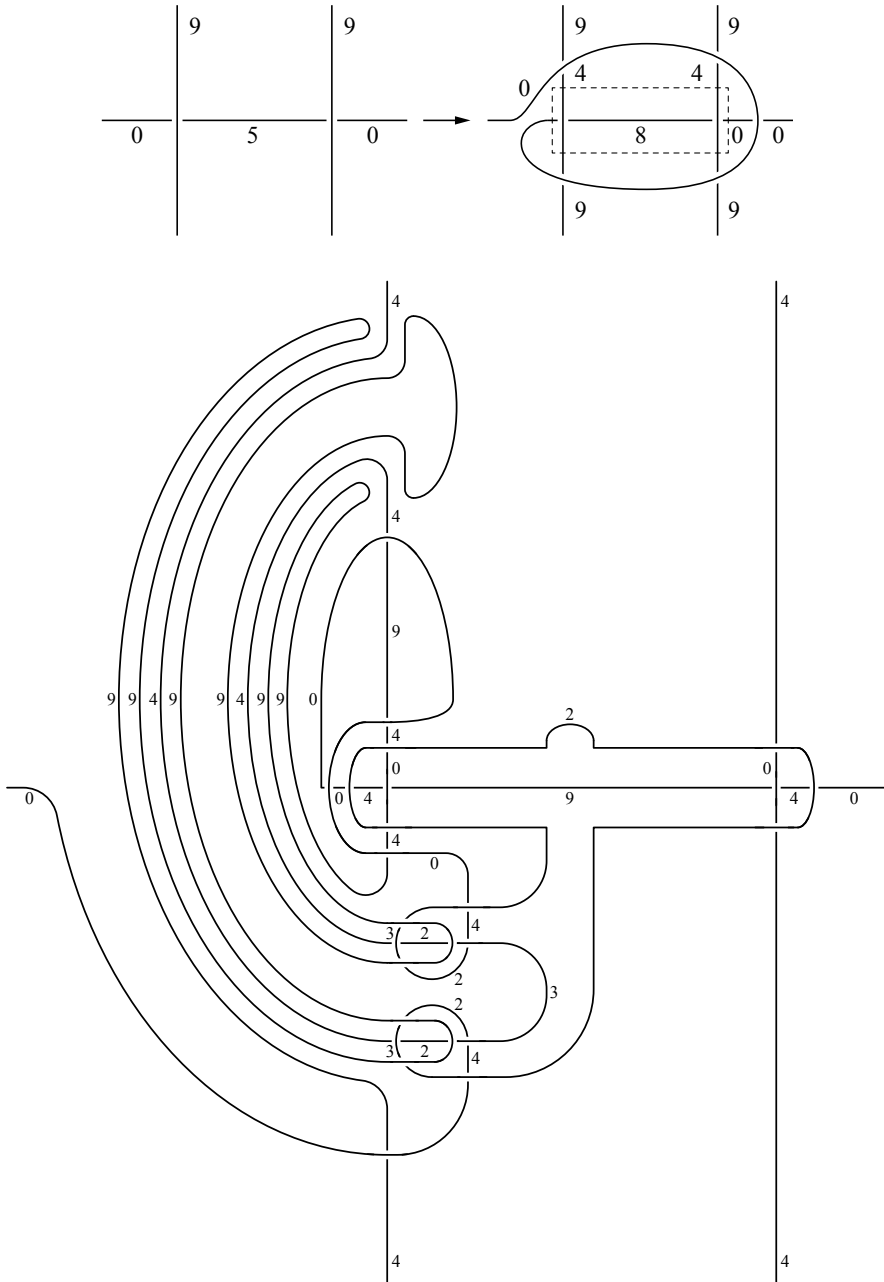


Figure 43

Next, we remove 5 as an over-arc by transforming any crossings of the form $\{a|5|10-a\}$. Since $10 - a \neq 1, 5, 6, 7, 8, 10, 11, 12$, we have $a \neq 0, 2, 3, 4, 9$. Therefore, 5 cannot be an over-arc.

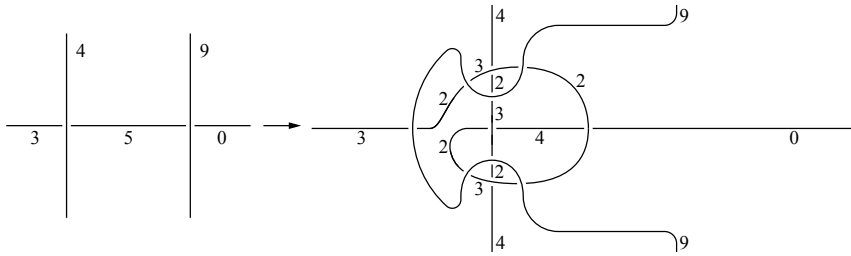


Figure 44

We complete the proof of Lemma 2.8 by removing 5 as an under-arc in a case-by-case method. We first consider the case where 5 is an under-arc connecting two crossings of the form $\{5|a|2a+8\}$. Since $2a+8 \neq 1, 5, 6, 7, 8, 10, 11, 12$, we have $a \neq 0, 2, 3$. So, we need to check $a = 4, 9$. If $a = 4$, we transform the diagram as in Figure 42. If $a = 9$, we transform the diagram as in Figure 43.

Now we consider the case where 5 is an under-arc connecting two crossings of the forms $\{5|a|2a+8\}$ and $\{5|b|2b+8\}$. Since $2a+8, 2b+8 \neq 1, 5, 6, 7, 8, 10, 11, 12$, there are two cases that we need to consider: $(a, b) = (4, 9), (9, 4)$. If $(a, b) = (4, 9)$, we transform the diagram as in Figure 44. The case $(a, b) = (9, 4)$ is similar. \square

At the same time we were working on this problem, Bento and Lopes [2015] proved the same result using different techniques.

References

- [Bento and Lopes 2015] F. Bento and P. Lopes, “The minimum number of Fox colors modulo 13 is 5”, preprint, 2015. arXiv
- [Fox 1962] R. H. Fox, “A quick trip through knot theory”, pp. 120–167 in *Topology of 3-manifolds and related topics* (Athens, GA, 1961), Prentice-Hall, Englewood Cliffs, N.J., 1962. MR Zbl
- [Hayashi et al. 2012] C. Hayashi, M. Hayashi, and K. Oshiro, “On linear n -colorings for knots”, *J. Knot Theory Ramifications* **21**:14 (2012), art. ID #1250123. MR Zbl
- [Lopes and Matias 2012] P. Lopes and J. Matias, “Minimum number of Fox colors for small primes”, *J. Knot Theory Ramifications* **21**:3 (2012), art ID #1250025. MR Zbl
- [Nakamura et al. 2013] T. Nakamura, Y. Nakanishi, and S. Satoh, “The pallet graph of a Fox coloring”, *Yokohama Math. J.* **59** (2013), 91–97. MR Zbl
- [Oshiro 2010] K. Oshiro, “Any 7-colorable knot can be colored by four colors”, *J. Math. Soc. Japan* **62**:3 (2010), 963–973. MR Zbl
- [Satoh 2009] S. Satoh, “5-colored knot diagram with four colors”, *Osaka J. Math.* **46**:4 (2009), 939–948. MR Zbl

Received: 2015-09-29

Revised: 2016-01-05

Accepted: 2016-01-24

emohamed@usf.edu

*Mathematics Department, University of South Florida,
Tampa, FL 33620, United States*

kerrj@mail.usf.edu

*Mathematics Department, University of South Florida,
Tampa, FL 33620, United States*

Combinatorial curve neighborhoods for the affine flag manifold of type A_1^1

Leonardo C. Mihalcea and Trevor Norton

(Communicated by Jim Haglund)

Let X be the affine flag manifold of Lie type A_1^1 . Its moment graph encodes the torus fixed points (which are elements of the infinite dihedral group D_∞) and the torus stable curves in X . Given a fixed point $u \in D_\infty$ and a degree $\mathbf{d} = (d_0, d_1) \in \mathbb{Z}_{\geq 0}^2$, the combinatorial curve neighborhood is the set of maximal elements in the moment graph of X which can be reached from u using a chain of curves of total degree $\leq \mathbf{d}$. In this paper we give a formula for these elements, using combinatorics of the affine root system of type A_1^1 .

1. Introduction

Let X be an arbitrary algebraic variety and $\Omega \subset X$ be a subvariety. Fix a degree \mathbf{d} , i.e., an effective homology class in $H_2(X)$. The (geometric) *curve neighborhood* $\Gamma_{\mathbf{d}}(\Omega)$ is the locus of points $x \in X$ which can be reached from Ω by a rational curve of some effective degree $\leq \mathbf{d}$. For example, if $X = \mathbb{P}^2$ is the projective plane, and $\Omega = \text{pt}$, then any other point in X can be reached from the given point, using a projective line. This implies $\Gamma_1(\text{pt}) = \mathbb{P}^2$.

Curve neighborhoods have been recently defined by A. Buch and the first author in [Buch and Mihalcea 2015] in relation to the study of quantum cohomology and quantum K theory rings of generalized flag manifolds $X = G/B$, where G is a complex semisimple Lie group and B is a Borel subgroup. The curve neighborhoods which are relevant in that context are those when Ω is a Schubert variety in G/B . It turns out that in this situation the calculation of the curve neighborhoods is encoded in the *moment graph* of X . This is a graph encoding the T -fixed points and the T -stable curves in X , where T is a maximal torus of G . Similar considerations, but in the case when X is an *affine* flag manifold, led L. Mare and the first author to a definition of an affine version of the quantum cohomology ring; see [Mare

MSC2010: primary 05E15; secondary 17B67, 14M15.

Keywords: affine flag manifolds, moment graph, curve neighborhood.

L. C. Mihalcea was supported in part by NSA Young Investigator Awards H98230-13-1-0208 and H98230-16-1-0013 and a Simons Collaboration Grant.

and Mihalcea 2014]. The curve neighborhoods which were relevant for quantum cohomology calculations were those for certain “small” degrees. Those for “large” degrees, which seem to encode more refined information about the geometry and the combinatorics of affine flag manifolds, are still unknown.

In the current paper we give an explicit combinatorial formula for the curve neighborhoods of the simplest affine flag manifold, that of affine Lie type A_1^1 . See, e.g., [Kumar 2002] for details on affine flag manifolds. Instead of introducing the geometry related to this flag manifold, we consider the more elementary—but equivalent—problem of calculating the *combinatorial* curve neighborhoods. These are encoded in the combinatorics of the moment graph of the affine flag manifold.

To state our main result, we briefly introduce some notation and recall a few definitions. Full details are given in Section 2 below. Let D_∞ be the infinite dihedral group, generated by reflections s_0 and s_1 . (This is the affine Weyl group of Lie type A_1^1 .) Each element of D_∞ has a unique reduced expression which involves a s_0 's and b s_1 's, where $|a - b| \leq 1$. There is a natural length function $\ell : D_\infty \rightarrow \mathbb{Z}_{\geq 0}$, and a (Bruhat) partial order on D_∞ , denoted $<$. A *degree* \mathbf{d} is a pair of nonnegative integers (d_0, d_1) . The moment graph has vertices given by the elements of D_∞ ; there is an edge between $u, v \in D_\infty$ whenever there exists an (affine root) reflection $s_{(a,b)}$ such that $v = us_{(a,b)}$. This edge has degree $\mathbf{d} = (a, b)$ such that $|a - b| = 1$; see Section 2B below. A *chain* in the moment graph is a succession of adjacent edges, and its degree is equal to the sum of the degrees of each of its edges.

Finally, fix a degree $\mathbf{d} = (d_0, d_1)$ and $u \in D_\infty$. The (combinatorial) *curve neighborhood* $\Gamma_{\mathbf{d}}(u)$ is the set of elements in D_∞ such that (1) they can be joined to u (in the moment graph) by a chain of degree $\leq \mathbf{d}$, and (2) they are maximal among all elements satisfying (1). To each $u \in D_\infty$, one associates the degree $d(u) := (a, b)$, where u has a reduced expression with a s_0 's and b s_1 's.

Consider the set

$$\mathcal{A}_{\mathbf{d}}(u) := \{v \in D_\infty : \ell(uv) = \ell(u) + \ell(v), d(v) \leq \mathbf{d}\},$$

and denote by $\max \mathcal{A}_{\mathbf{d}}(u)$ the subset of its maximal elements. Our main result is:

Theorem 1.1. *Let $u \in D_\infty$ and $\mathbf{d} = (d_1, d_2)$ be a degree. Then the following hold:*

(a) *The curve neighborhood $\Gamma_{\mathbf{d}}(u)$ is given by*

$$\Gamma_{\mathbf{d}}(u) = \{uw : w \in \max \mathcal{A}_{\mathbf{d}}(u)\}.$$

(b) *Formulas (3) and (4) below give explicit combinatorial formulas for the elements in $\max \mathcal{A}_{\mathbf{d}}(u)$. In particular, the curve neighborhood $\Gamma_{\mathbf{d}}(u)$ has exactly two elements if $u = 1$ and $\mathbf{d} = (a, a)$, and one element otherwise.*

It is interesting to remark that the curve neighborhoods distinguish the degrees corresponding to “imaginary roots” (a, a) in this case. (See [Kac 1985] for more

about this affine root system.) We plan to study further this phenomenon elsewhere. The theorem implies the “geometric” curve neighborhood for the Schubert variety indexed by u is either a single Schubert variety, or the union of two Schubert varieties, indexed by the elements in $\Gamma_d(u)$. We refer to [Mare and Mihalcea 2014] for a discussion of geometric curve neighborhoods.

This paper is the outcome of an undergraduate research project of Norton conducted under the direction of Mihalcea.

2. Preliminaries

2A. The infinite dihedral group. The infinite dihedral group D_∞ is the group with generators s_0, s_1 and relations $s_0^2 = s_1^2 = 1$. Each element $w \in D_\infty$ can be written *uniquely* as a product of s_0 's and s_1 's in such a way that no s_0 's and no s_1 's are consecutive. We call such an expression *reduced*. We define the *length* $\ell(w)$ of w to be the total number of s_0 's and s_1 's in the expression of w . For example, $\ell(s_0) = 1$ and $\ell(s_1s_0s_1s_0) = 4$.

A (positive) *root* corresponding to D_∞ is a pair of nonnegative integers $\alpha = (a, b) \in \mathbb{Z}_{\geq 0}^2$ such that $|a - b| = 1$. For example, $\alpha_0 := (1, 0)$ and $\alpha_1 := (0, 1)$ are roots, and so is $\alpha = 2\alpha_0 + 3\alpha_1 = (2, 3)$. Fix a root $\alpha = (a_0, a_1)$. A *root reflection* s_α is the unique element of D_∞ which can be written as a product of a_0 s_0 's and a_1 s_1 's, and which has length $a + b$. For example,

$$s_{(2,3)} = s_1s_0s_1s_0s_1, \quad s_{(1,0)} = s_0.$$

The terminology follows from the fact that these are the positive roots of the affine Lie algebra of type A_1^1 ; see, e.g., [Kac 1985].

We record for later use the following properties:

Lemma 2.1. *Let $u, v \in D_\infty$. Then:*

- (a) $\ell(u) = \ell(u^{-1})$.
- (b) u is a root reflection if and only if $\ell(u)$ is odd.
- (c) $\ell(uv) \leq \ell(u) + \ell(v)$.
- (d) If $\ell(u) \leq \ell(v)$, then

$$\ell(uv) = \ell(u) + \ell(v) \quad \text{or} \quad \ell(uv) = \ell(v) - \ell(u).$$

In particular, $\ell(uv) \equiv \ell(u) + \ell(v) \pmod{2}$.

Proof. This is an easy verification. □

2B. The moment graph and curve neighborhoods. The *moment graph* G associated to D_∞ is the graph given by the following data:

- The set V of *vertices* is the group D_∞ .

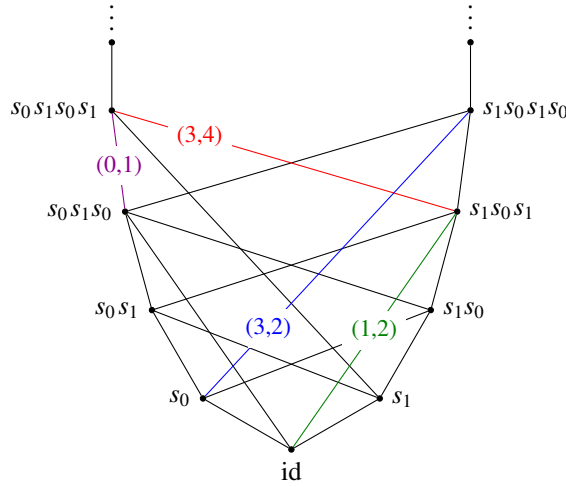


Figure 1. The moment graph G associated to D_∞ .

- Let $u, v \in V$ be vertices. Then there is an edge from u to v if and only if there exists a root $\alpha = (a_0, a_1)$ such that $v = us_\alpha$. We denote this situation by

$$u \xrightarrow{\alpha} v,$$

and we say that the *degree* of this edge is α .

In Figure 1 we show the moment graph up to elements of length 4. We labeled a few of the edges by their corresponding degrees.

Remark 2.2. As mentioned in the Introduction, the vertices of this graph correspond to the T -fixed points, and its edges to the T -stable curves in the affine flag manifold of type A_1^1 , where T is a maximal torus in an affine Kac–Moody group of type A_1^1 . See, e.g., [Kumar 2002, Chapter 12], especially §12.2.E, for details.

A *chain* between u and v in the moment graph is a succession of adjacent edges starting with u and ending with v :

$$\pi : u = u_0 \xrightarrow{\beta_0} u_1 \xrightarrow{\beta_1} \dots \xrightarrow{\beta_{n-2}} u_{n-1} \xrightarrow{\beta_{n-1}} u_n = v.$$

The chain is called *increasing* if at each step the lengths increase, i.e., $\ell(u_i) > \ell(u_{i-1})$ for $1 \leq i \leq n$. The *degree* of the chain π is $\deg(\pi) = \beta_0 + \dots + \beta_{n-1}$. Define a partial ordering on the elements of D_∞ by $u < v$ if and only if there exists an increasing chain starting with u and ending with v .

The next result gives an equivalent way to describe the partial ordering on D_∞ :

Lemma 2.3. *Let $u, v \in D_\infty$. Then $u < v$ if and only if $\ell(u) < \ell(v)$.*

Proof. Clearly if $u < v$ then $\ell(u) < \ell(v)$ from the definition of an increasing chain. To prove the converse, we first notice that if $\ell(v) - \ell(u) = 1$, then $u^{-1}v$ is a root reflection s_α (possibly of length > 1); thus there exists an edge $u \xrightarrow{\alpha} v$. The general statement follows by induction on $\ell(v) - \ell(u) \geq 1$. \square

A *degree* is a pair of nonnegative integers $\mathbf{d} = (d_0, d_1)$. There is a natural partial order on degrees. If $\mathbf{d} = (d_0, d_1)$ and $\mathbf{d}' = (d'_0, d'_1)$ then $\mathbf{d} \geq \mathbf{d}'$ if and only if $d_i \geq d'_i$ for $i \in \{0, 1\}$.

Definition 2.4. Fix a degree \mathbf{d} and $u \in D_\infty$. The (*combinatorial*) *curve neighborhood* is the set $\Gamma_{\mathbf{d}}(u)$ consisting of elements $v \in D_\infty$ such that

- (1) there exists a chain of some degree $\mathbf{d}' \leq \mathbf{d}$ from u to v in the moment graph G ;
- (2) the elements v are maximal among all of those satisfying the condition in (1).

For example,

$$\Gamma_{(1,0)}(\text{id}) = \Gamma_{(2,0)}(\text{id}) = \{s_0\}, \quad \Gamma_{(1,1)}(\text{id}) = \{s_1s_0, s_0s_1\}.$$

Our main goal is to find a formula to determine $\Gamma_{\mathbf{d}}(u)$.

For $w \in D_\infty$, define the degree associated to w to be $d(w) = (d_0, d_1)$, where $d_i :=$ number of reflections s_i in the reduced word of w . The following holds.

Lemma 2.5. *Let $u, v \in D_\infty$ and assume there is a chain from u to v of degree \mathbf{d} . Then $\mathbf{d} = d(u^{-1}v) + 2(r, s)$, where $r, s \in \mathbb{Z}_{\geq 0}$. In particular, $\mathbf{d} \geq d(u^{-1}v)$.*

Proof. Let $\beta_0 := (a_0, b_0)$, $\beta_1 := (a_1, b_1)$, \dots , $\beta_{n-1} := (a_{n-1}, b_{n-1})$ be the labels of the edges of the chain π . Then $v = us_{\beta_0} \cdots s_{\beta_{n-1}}$ and $\mathbf{d} = \beta_0 + \cdots + \beta_{n-1} = (a_0 + \cdots + a_{n-1}, b_0 + \cdots + b_{n-1})$. Now $d(u^{-1}v) = d(s_{\beta_0} \cdots s_{\beta_{n-1}})$. If $s_{\beta_0} \cdots s_{\beta_{n-1}}$ is nonreduced, one needs to perform some cancellations of the form $s_0^2 = 1$ or $s_1^2 = 1$. Each of these result in a decrease by 2 of the number of s_0 's, respectively s_1 's, in an expression for $s_{\beta_0} \cdots s_{\beta_{n-1}}$. Thus $d(u^{-1}v) = \mathbf{d} - 2(r, s)$ as claimed. \square

3. Calculation of the curve neighborhoods

Let $\mathbf{d} = (d_1, d_2)$ be a degree such that $d_1 \neq d_2$. We denote by $\alpha(\mathbf{d})$ the maximal root α such that $\alpha \leq \mathbf{d}$. Clearly there is exactly one such root, and it is easy to find the following explicit formula for it:

$$\alpha(\mathbf{d}) = \begin{cases} (d_1, d_1 + 1) & \text{if } d_1 < d_2, \\ (d_2 + 1, d_2) & \text{if } d_1 > d_2. \end{cases} \tag{1}$$

In order to find the curve neighborhoods of an element $u \in D_\infty$, we need the following key result.

Lemma 3.1. *Let $u \in D_\infty$ and $\mathbf{d} = (d_1, d_2)$ be a degree. Consider the set*

$$\mathcal{A}_{\mathbf{d}}(u) := \{v \in D_\infty : \ell(uv) = \ell(u) + \ell(v), d(v) \leq \mathbf{d}\}. \tag{2}$$

Then the following hold:

- $\mathcal{A}_d(u)$ has a unique maximal element if $u \neq 1$ or if $u = 1$ and $d \neq (a, a)$ for some nonnegative integer a .
- If $d = (a, a)$ and $u = 1$ then the maximal elements of $\mathcal{A}_d(u)$ are $(s_0s_1)^a$ and $(s_1s_0)^a$.

Proof. Clearly, $1 \in \mathcal{A}_d(u)$ so $\mathcal{A}_d(u) \neq \emptyset$. For any $v \in \mathcal{A}_d(u)$, we have $d(v) \leq d$. Hence the set $\mathcal{A}_d(u)$ is finite, and so it must contain a maximal element. Lemma 2.3 implies there can be at most two maximal elements v_1 and v_2 and they must have the same length. We consider each of the situations in the statement:

Case 1: $u \neq 1$. Assume there are two maximal elements v_1, v_2 . Since $u \neq 1$, either uv_1 or uv_2 is not reduced, say uv_1 . Then $\ell(uv_1) < \ell(u) + \ell(v_1)$, and this contradicts that $v_1 \in \mathcal{A}_d(u)$.

Case 2: $u = 1$. In this case, the set $\mathcal{A}_d(u)$ coincides with the set of all $v \in D_\infty$ such that $d(v) \leq d$. From the description of D_∞ , it follows that $d(v) = (a, a)$ or $d(v) = (a, a+1)$ or $d(v) = (a+1, a)$ for some nonnegative integer a . Further, the reduced decomposition of v is known in each case: there are two possibilities for v if $d(v) = (a, a)$, and there is exactly one (in fact, $v = s_{\alpha(d)}$) in the other two cases. The claim follows from this. \square

In what follows, we will denote by $\max \mathcal{A}_d(u)$ the set of maximal elements in the (finite) partially ordered set $\mathcal{A}_d(u)$. Our main result is:

Theorem 3.2. *Let $u \in D_\infty$ and $d = (d_1, d_2)$ be a degree. Then*

$$\Gamma_d(u) = \{uw : w \in \max \mathcal{A}_d(u)\}.$$

We will prove this theorem in the next two sections, which correspond to the cases $u = 1$ and $u \neq 1$. For now, notice that the proof of Lemma 3.1, and some easy arguments based on reduced decompositions in D_∞ , imply that if $u = 1$ then the set of maximal elements of $\mathcal{A}_d(1)$ is

$$\max \mathcal{A}_d(1) = \begin{cases} \{s_{\alpha(d)}\} & \text{if } d = (d_1, d_2) \text{ and } d_1 \neq d_2, \\ \{(s_0s_1)^a, (s_1s_0)^a\} & \text{if } d = (a, a). \end{cases} \quad (3)$$

If $u \neq 1$, we assume for simplicity that last simple reflection in the reduced word for u is s_0 , i.e., $u = \cdots s_0$. (The other situation will be symmetric). Then

$$\max \mathcal{A}_d(u) = \begin{cases} \{s_1s_{\alpha(d-(0,1))}\} & \text{if } d_0 = d_1, \\ \{s_{\alpha(d)}\} & \text{if } d_1 > d_0, \\ \{s_0s_{\alpha(d)}\} & \text{if } d_1 < d_0. \end{cases} \quad (4)$$

The two formulas give explicit combinatorial rules to determine the curve neighborhood $\Gamma_d(u)$. See Section 3C below for several examples.

3A. Curve neighborhoods for $u = 1$.

Theorem 3.3. *Let $\mathbf{d} = (d_1, d_2)$ be a degree. Then the curve neighborhood of the identity can be calculated in the following way:*

$$\Gamma_{\mathbf{d}}(1) = \max \mathcal{A}_{\mathbf{d}}(1) = \begin{cases} \{s_{\alpha}(\mathbf{d})\} & \text{if } d_1 \neq d_2, \\ \{(s_0s_1)^a, (s_1s_0)^a\} & \text{if } \mathbf{d} = (a, a). \end{cases}$$

Proof. If $v \in \Gamma_{\mathbf{d}}(1)$ then there exists a chain of degree $\leq \mathbf{d}$ joining 1 to v . Then by Lemma 2.5, $\mathbf{d} \geq d(v)$. In particular, $v \in \mathcal{A}_{\mathbf{d}}(1)$; thus $\Gamma_{\mathbf{d}}(1) \subset \mathcal{A}_{\mathbf{d}}(1)$, and the inclusion is compatible with the partial order $<$. Conversely, if v is any element in $\mathcal{A}_{\mathbf{d}}(1)$ then there exists a chain of degree $d(v) \leq \mathbf{d}$ joining 1 to v . If v is maximal in $\mathcal{A}_{\mathbf{d}}(1)$, and because $\Gamma_{\mathbf{d}}(1) \subset \mathcal{A}_{\mathbf{d}}(1)$, it follows that $v \in \Gamma_{\mathbf{d}}(1)$. \square

3B. General curve neighborhoods. The goal of this section is to find a formula for the curve neighborhoods $\Gamma_{\mathbf{d}}(u)$ for $u \neq 1$ and $\mathbf{d} \neq (0, 0)$. First we need some preparatory lemmas.

Lemma 3.4. *Let $u \in D_{\infty}$, $z \in \mathcal{A}_{\mathbf{d}}(1)$ and $v \in \Gamma_{\mathbf{d}}(u)$. Then:*

- (a) $\ell(uz) \leq \ell(v)$ and $d(u^{-1}v) \leq \mathbf{d}$.
- (b) If $z \in \Gamma_{\mathbf{d}}(1)$ (i.e., z is maximal in $\mathcal{A}_{\mathbf{d}}(1)$), then $\ell(u^{-1}v) \leq \ell(z)$.

Proof. Since $z \in \mathcal{A}_{\mathbf{d}}(1)$, there exists a chain of degree $d(z) \leq \mathbf{d}$ joining 1 to z . Multiplying this chain by u on the left gives a chain between u and uz of the same degree. The first statement in (a) follows by the maximality of v . To prove the second statement in (a), notice that since $v \in \Gamma_{\mathbf{d}}(u)$, there exists a chain from u to v of degree $\leq \mathbf{d}$. If we multiply each element of this chain on the left by u^{-1} , we obtain a chain from 1 to $u^{-1}v$ of the same degree. The fact that $d(u^{-1}v) \leq \mathbf{d}$ follows from Lemma 2.5. Finally, (b) follows from the maximality of z , using also that maximal elements in $\mathcal{A}_{\mathbf{d}}(1)$ have the same length, by (3). \square

The following lemma gives a strong constraint on the possible elements in $\Gamma_{\mathbf{d}}(u)$.

Lemma 3.5. *Let $v \in \Gamma_{\mathbf{d}}(u)$. Then $u^{-1}v \in \mathcal{A}_{\mathbf{d}}(u)$.*

Proof. We have seen in Lemma 3.4 that $d(u^{-1}v) \leq \mathbf{d}$. It remains to show that $\ell(v) = \ell(u) + \ell(u^{-1}v)$. This clearly holds for $u = 1$ and from now on we assume $u \neq 1$. From Lemma 2.1(c) it follows that $\ell(v) = \ell(uu^{-1}v) \leq \ell(u) + \ell(u^{-1}v)$. If the inequality is strict then $\ell(u^{-1}v) > \ell(v) - \ell(u) = \ell(v) - \ell(u^{-1})$. But $\ell(u) \leq \ell(v)$, thus by Lemma 2.1(d) it follows that

$$\ell(u^{-1}v) = \ell(u) + \ell(v).$$

Consider now an element $z \in \Gamma_{\mathbf{d}}(1) = \max \mathcal{A}_{\mathbf{d}}(1)$ (by Theorem 3.3). We invoke Lemma 3.4 to obtain

$$\ell(uz) \leq \ell(v) < \ell(u) + \ell(u^{-1}v) \leq \ell(u) + \ell(z).$$

This implies the expression uz is not reduced. But since $u \neq 1$, we can eliminate the first simple reflection from the reduced expression for z to define $z' < z$ such that $\ell(z') = \ell(z) - 1$ and $\ell(uz') = \ell(u) + \ell(z')$. Notice that $d(z') < d(z) \leq \mathbf{d}$; thus $z' \in \mathcal{A}_{\mathbf{d}}(1)$. Then we have the inequalities

$$\ell(v) \geq \ell(uz') = \ell(u) + \ell(z) - 1 \geq \ell(u) + \ell(u^{-1}v) - 1 = \ell(u) + \ell(u) + \ell(v) - 1,$$

where the first inequality follows from Lemma 3.4(a) and the last inequality follows from Lemma 3.4(b). Taking the extreme sides and subtracting $\ell(v)$, we obtain $0 \geq 2\ell(u) - 1$, which is impossible since $\ell(u) \geq 1$. Thus $\ell(v) = \ell(u) + \ell(u^{-1}v)$ and this finishes the proof. \square

We are ready to prove our main result.

Theorem 3.6. *Let $\mathbf{d} = (d_1, d_2)$ be a nonzero degree and $u \in D_\infty$. Then*

$$\Gamma_{\mathbf{d}}(u) = \{uw : w \in \max \mathcal{A}_{\mathbf{d}}(u)\}.$$

Proof. Let $v \in \Gamma_{\mathbf{d}}(u)$. Then Lemma 3.5 implies $u^{-1}v \in \mathcal{A}_{\mathbf{d}}(u)$. From Lemma 3.1 (or (4)), there exists a unique maximal element of $\mathcal{A}_{\mathbf{d}}(u)$, call it w . Then $u^{-1}v \leq w$ and clearly w is also in $\mathcal{A}_{\mathbf{d}}(1)$. By Lemma 3.4(a), we deduce $\ell(uw) \leq \ell(v)$. Then

$$\ell(u) + \ell(u^{-1}v) = \ell(v) \geq \ell(uw) = \ell(u) + \ell(w).$$

This implies $\ell(u^{-1}v) \geq \ell(w)$. Together with $u^{-1}v \leq w$, this forces $u^{-1}v = w$; i.e., $v = uw$ as claimed. \square

3C. Examples. We provide several examples determining $\Gamma_{\mathbf{d}}(u)$.

- Let $u = 1$ and $\mathbf{d} = (9, 4)$. From (1) we obtain $\alpha(\mathbf{d}) = (5, 4)$; thus

$$\Gamma_{(9,4)}(1) = \{s_{(5,4)}\} = \{s_0s_1s_0s_1s_0s_1s_0s_1s_0\}.$$

- Let $u = 1$ and $\mathbf{d} = (4, 4)$. By (3) the two maximal elements in $\mathcal{A}_{(4,4)}(1)$ are $s_0s_1s_0s_1s_0s_1s_0s_1$ and $s_1s_0s_1s_0s_1s_0s_1s_0$. Then

$$\Gamma_{(4,4)}(\text{id}) = \{s_0s_1s_0s_1s_0s_1s_0s_1, s_1s_0s_1s_0s_1s_0s_1s_0\}.$$

- Let $u = s_0s_1s_0$ and $\mathbf{d} = (3, 3)$. From (4),

$$\max \mathcal{A}_{(3,3)}(u) = \{s_1s_{\alpha((3,3)-(0,1))}\} = \{s_1s_0s_1s_0s_1s_0\}.$$

Thus $\Gamma_{(3,3)}(s_0s_1s_0) = \{(s_0s_1s_0)(s_1s_0s_1s_0s_1s_0)\}$.

- Let $u = s_1s_0s_1$ and $\mathbf{d} = (3, 3)$. From the symmetric version of (4),

$$\max \mathcal{A}_{(3,3)}(u) = \{s_0s_{\alpha((3,3)-(1,0))}\} = \{s_0s_1s_0s_1s_0s_1\}.$$

Thus $\Gamma_{(3,3)}(s_1s_0s_1) = \{(s_1s_0s_1)(s_0s_1s_0s_1s_0s_1)\}$.

- Let $u = s_0 s_1 s_0$ and $d = (9, 4)$. Then $\alpha((9, 4)) = (5, 4)$ and using (4) again, $\max \mathcal{A}_d(u) = \{s_1 s_0 s_1 s_0 s_1 s_0 s_1 s_0\}$. Then

$$\Gamma_{(9,4)}(s_0 s_1 s_0) = \{(s_0 s_1 s_0)(s_1 s_0 s_1 s_0 s_1 s_0 s_1 s_0)\}.$$

- Let $u = s_0 s_1 s_0$ and $d = (4, 9)$. Then $\alpha((4, 9)) = (4, 5)$ and $\max \mathcal{A}_d(u) = \{s_1 s_0 s_1 s_0 s_1 s_0 s_1 s_0 s_1\}$. From this we obtain

$$\Gamma_{(9,4)}(s_0 s_1 s_0) = \{(s_0 s_1 s_0)(s_1 s_0 s_1 s_0 s_1 s_0 s_1 s_0 s_1)\}.$$

References

- [Buch and Mihalcea 2015] A. S. Buch and L. C. Mihalcea, “Curve neighborhoods of Schubert varieties”, *J. Differential Geom.* **99**:2 (2015), 255–283. MR Zbl
- [Kac 1985] V. G. Kac, *Infinite-dimensional Lie algebras*, 2nd ed., Cambridge University Press, 1985. MR Zbl
- [Kumar 2002] S. Kumar, *Kac–Moody groups, their flag varieties and representation theory*, Progress in Mathematics **204**, Birkhäuser, Boston, 2002. MR Zbl
- [Mare and Mihalcea 2014] L. Mare and L. C. Mihalcea, “An affine quantum cohomology ring for flag manifolds and the periodic Toda lattice”, preprint, 2014. arXiv

Received: 2015-12-13 Accepted: 2016-04-01

lmihalce@math.vt.edu

*Department of Mathematics, Virginia Tech University,
Blacksburg, VA 24061, United States*

norton15@vt.edu

*Department of Mathematics, Virginia Tech University,
Blacksburg, VA 24061, United States*

Total variation based denoising methods for speckle noise images

Arundhati Bagchi Misra, Ethan Lockhart and Hyeona Lim

(Communicated by Kenneth S. Berenhaut)

In this paper, we introduce a new algorithm based on total variation for denoising speckle noise images. Total variation was introduced by Rudin, Osher, and Fatemi in 1992 for regularizing images. Chambolle proposed a faster algorithm based on the duality of convex functions for minimizing the total variation, but his algorithm was built for Gaussian noise removal. Unlike Gaussian noise, which is additive, speckle noise is multiplicative. We modify the original Chambolle algorithm for speckle noise images using the first noise equation for speckle denoising, proposed by Krissian, Kikinis, Westin and Vosburgh in 2005. We apply the Chambolle algorithm to the Krissian et al. speckle denoising model to develop a faster algorithm for speckle noise images.

1. Introduction

Image restoration, especially image denoising, is a very important process and is often necessary as preprocessing for other imaging techniques such as segmentation and compression. For the last two decades, various partial differential equation (PDE) based models have been developed for this purpose [Rudin et al. 1992; Perona and Malik 1990; Kornprobst et al. 1997; Catté et al. 1992; Alvarez et al. 1992; Chan and Vese 1997; Vese and Chan 1997; Marquina and Osher 2000; Chan et al. 1999; Joo and Kim 2003a; 2003b; Kim 2004; Kim and Lim 2007]. In general, an observed image f , corrupted by Gaussian noise n , is represented by the equation

$$f = u + n, \tag{1}$$

where u is the original noise-free image. Here $u, f : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}$. For any denoising model, the main objective is to reconstruct u from an observed image f .

MSC2010: primary 68U10, 94A08; secondary 65M06, 65N06, 65K10, 49K20.

Keywords: anisotropic diffusion, speckle noise, denoising, total variation (TV) model, Chambolle algorithm, fast speckle denoising.

Rudin, Osher, and Fatemi [Rudin et al. 1992] proposed the total variation (TV) denoising model as the minimization problem

$$\min_u \int_{\Omega} |\nabla u| \, d\vec{x} \quad (2)$$

subject to the constraints,

$$\int_{\Omega} f \, d\vec{x} = \int_{\Omega} u \, d\vec{x}, \quad (3)$$

$$\int_{\Omega} \frac{1}{2}(f - u)^2 \, d\vec{x} = \sigma^2, \quad (4)$$

where σ is the standard deviation of the noise n . These constraints ensure that the resulting image and the observed image are close to each other.

Combining the above constraints, the TV functional is obtained by

$$F(u) = \int_{\Omega} |\nabla u| \, d\vec{x} + \frac{\lambda}{2} \int_{\Omega} (f - u)^2 \, d\vec{x}. \quad (5)$$

Here, λ is a constraint parameter. The equivalent Euler–Lagrange equation gives the TV denoising model as

$$\frac{\partial u}{\partial t} - \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right) = \lambda(f - u). \quad (6)$$

To avoid singularities, it was regularized by using $|\nabla u| \approx |\nabla^\varepsilon u| = (u_x^2 + u_y^2 + \varepsilon^2)^{1/2}$.

In this paper, we introduce a faster denoising method, compared to TV model, for speckle noise images. Unlike Gaussian noise, speckle noise is multiplicative and requires a model separate from those for Gaussian noise images. The first effective speckle denoising model was developed by Krissian, Kikinis, Westin, and Vosburgh [Krissian et al. 2005], and they proposed a new noise equation for speckle denoising. For our new model, we modify the original Chambolle algorithm designed for the TV model (6), and develop a fast and accurate speckle denoising method based on the noise equation proposed by Krissian et al.

2. The Chambolle algorithm: dual approach

In this section, we provide a brief description of the Chambolle algorithm for the TV model. Chambolle [2004] provided a fast algorithm for minimizing the total variation. Detailed background and development of the algorithm can be found in his lecture notes [Chambolle et al. 2010]. The work is based on the dual formulation of Chan, Golub, and Mulet [Chan et al. 1999] and of Carter [2001]. To avoid the staircasing effect, he derived the Euler–Lagrange equation in the sense of convex analysis. His paper also contains the proof of the convergence of his algorithm.

The Chambolle algorithm. Chambolle [2004] started with the Rudin, Osher, and Fatemi (ROF) minimization functional [Rudin et al. 1992]

$$\min_u \left[\lambda J(u) + \int_{\Omega} \frac{1}{2} |u - f|^2 d\vec{x} \right], \tag{7}$$

where $J(u) = \int_{\Omega} |\nabla u| d\vec{x}$. He proved that u is a minimizer of (7) if and only if

$$\frac{f - u}{\lambda} \in \partial J(u),$$

where ∂F denotes the subdifferential of a convex function F . Hence, the Euler–Lagrange equation obtained by Chambolle is given as

$$\lambda \partial J(u) + u - f \ni 0. \tag{8}$$

Note that $u = (u_{ij})$, where $i, j = 1, \dots, N$, is the discrete image. Thus $u \in X = \mathbb{R}^{N \times N}$. Applying the Legendre–Fenchel identity, he obtained the dual problem as

$$\min_{|p_{i,j}| \leq 1} \frac{1}{2} \left\| \operatorname{div} p - \frac{f}{\lambda} \right\|^2, \tag{9}$$

where $\operatorname{div} p = w$ for $p = (p_{i,j})$ with $i, j = 1, \dots, N \in Y = X \times X$ and $w = (f - u)/\lambda$. Here $\| \cdot \|$ is the Euclidean norm, which is defined similarly to (6) in [Chambolle 2004]. We can recover u by

$$u = f - \lambda w = f - \lambda \operatorname{div} p. \tag{10}$$

Hence, to find the denoised image u , the following problem must be solved for p :

$$\min \{ \| \lambda \operatorname{div} p - f \|^2 : p \in Y, |p_{i,j}| \leq 1 \forall i, j = 1, \dots, N \}. \tag{11}$$

Chambolle proposed the following algorithm to solve for p . Choosing $\tau > 0$ and taking $p^0 = 0$ we derive p^n for any $n > 0$, by

$$p_{i,j}^{n+1} = \frac{p_{i,j}^n + \tau (\nabla (\operatorname{div} p^n - f/\lambda))_{i,j}}{1 + \tau |(\nabla (\operatorname{div} p^n - f/\lambda))_{i,j}|}. \tag{12}$$

The next theorem proved the convergence of the algorithm in [Chambolle 2004]:

Theorem 2.1. *If $\tau < \frac{1}{8}$, then $f - \lambda \operatorname{div} p^n$ converges to u as $n \rightarrow \infty$.*

Image denoising with the Chambolle algorithm. In general, for Gaussian noise images, Chambolle proposed updating λ at each iteration by the formula

$$\lambda^{n+1} = \frac{N\sigma}{g_n} \lambda^n = \frac{N\sigma}{\| \operatorname{div} p^{n+1} \|} = \frac{\| f - u_c \|}{\| \operatorname{div} p^{n+1} \|}, \tag{13}$$

where u_c is the noise-free clear image and the variance σ^2 of the noise is known. For $s > 0$, he defined $g(s) = \| s \operatorname{div} p \|$. Here, the starting value $\lambda_0 > 0$ is chosen

arbitrarily. Thus the Chambolle algorithm for TV denoising [Chambolle 2004; Chambolle et al. 2010] is given by

$$\begin{aligned} p_{i,j}^{n+1} &= \frac{p_{i,j}^n + \tau(\nabla(\operatorname{div} p^n - f/\lambda))_{i,j}}{1 + \tau|(\nabla(\operatorname{div} p^n - f/\lambda))_{i,j}|}, \\ \lambda^{n+1} &= \frac{\|f - u_c\|}{\|\operatorname{div} p^{n+1}\|}, \\ u^{n+1} &= f - \lambda^{n+1} \operatorname{div} p^{n+1} \end{aligned}$$

for any $n \geq 0$. This algorithm converges almost twice as fast as the regular TV model (6).

3. Speckle noise

Speckle noise is mostly present in ultrasound images, synthetic aperture radar (SAR) images, and acoustic images. It is granular in nature, and it exists inherently in the image. Unlike Gaussian noise, which affects single pixels of an image, speckle noise affects multiple pixels. The noise is multiplicative, whereas Gaussian noise is additive. Hence, it is not possible to remove speckle noise with the traditional Gaussian denoising models.

Speckle denoising model by Krissian et al. This model was proposed by Krissian, Kikinis, Westin, and Vosburgh [Krissian et al. 2005], where they mainly dealt with speckle noise present in ultrasound images. They considered the speckle noise equation as

$$f = u + \sqrt{un}, \quad (14)$$

where u is the desired image to find, n is Gaussian noise, and f is the observed image. Hence using $n = (f - u)/\sqrt{u}$, the general regularized minimization functional is given as

$$\min_u F(u) = \min_u \left(\int_{\Omega} \left[|\nabla u| + \frac{\lambda}{2} \left(\frac{f - u}{\sqrt{u}} \right)^2 \right] d\bar{x} \right).$$

Finally using the Euler–Lagrange equation of this functional, the TV based speckle denoising model [Marquina and Osher 2000; Kim and Lim 2007] is derived as

$$\frac{\partial u}{\partial t} - \frac{u^2}{f + u} |\nabla u| \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right) = \lambda |\nabla u| (f - u). \quad (15)$$

4. The modified Chambolle for speckle denoising (MCSD) model

The Chambolle algorithm gives faster results than the regular TV model. Unfortunately, the model is formulated to work only for synthetic Gaussian noises.

We modify it to obtain the modified Chambolle for speckle denoising (MCSD) model for natural speckle noise. In [Wen et al. 2016] the authors used the primal-dual approach of Chambolle [Chambolle and Pock 2011] to develop a TV based denoising model for Poisson noise images. Similar approaches were proposed for multiplicative-noise images based on the Chambolle primal-dual algorithm in [Chan et al. 2014; Huang et al. 2012; 2013a; 2013b; Dong and Zeng 2013] for image segmentation, denoising and deblurring.

The MCSD model is based on the Chambolle algorithm for faster TV denoising. We apply the Chambolle-TV algorithm on the Krissian et al. speckle model to obtain a faster speckle denoising model. We start with developing the Euler–Lagrange equation for MCSD and then build the algorithm based on it.

The Euler–Lagrange equation for MCSD. We start by developing the Euler–Lagrange equation based on the speckle noise equation (14) introduced by Krissian et al. The minimization functional will be given by

$$\min_u \left[\lambda J(u) + \int_{\Omega} \frac{1}{2} \frac{|u - f|^2}{u} d\vec{x} \right]. \quad (16)$$

A similar model has also been discussed in [Jin and Yang 2011]. In this paper, the authors also developed the denoising functional (16) for speckle noise images motivated by the ROF model [Rudin et al. 1992] and the speckle noise model by Krissian et al. [2005]. They proved the existence and uniqueness of the minimizer for the functional (16). The existence and uniqueness of weak solutions for the associated evolution equation were also derived. For numerical computation, they directly used the finite difference scheme for the evolution equation, based on the schemes introduced in [Rudin et al. 1992]. However, in our paper, we adopt a dual formation suggested by Chambolle [2004] (see also [Chambolle et al. 2010]) for (16) to produce a faster algorithm. This is the main difference between the results of [Jin and Yang 2011] and ours. Moreover, for the purpose of development of the Euler–Lagrange equation, we consider the following slightly modified functional by using the fact $u \approx f$:

$$\min_u \left[\lambda J(u) + \int_{\Omega} \frac{1}{2} \frac{|u - f|^2}{f} d\vec{x} \right]. \quad (17)$$

The following Theorem 4.1 provides us with the formulation of the Euler–Lagrange equation for our new model.

Theorem 4.1. *The Euler–Lagrange equation for the minimizing functional (16) is*

$$\partial J(u) + \frac{u - f}{\lambda u} \ni 0. \quad (18)$$

Proof. If u is the solution of the functional (16), since $u \approx f$, we have for any $v \in L^2(\Omega)$,

$$\begin{aligned} \lambda J(v) + \frac{1}{2} \int_{\Omega} \frac{|v-f|^2}{f} \, d\vec{x} &\geq \lambda J(u) + \frac{1}{2} \int_{\Omega} \frac{|u-f|^2}{f} \, d\vec{x} \\ &\Rightarrow \lambda J(v) \geq \lambda J(u) + \frac{1}{2} \int_{\Omega} \frac{1}{f} [(u-f)^2 - (v-f)^2] \, d\vec{x} \\ &= \lambda J(u) + \int_{\Omega} \frac{u-v}{f} \left(\frac{v-u}{2} - (f-u) \right) \, d\vec{x} \\ &= \lambda J(u) - \int_{\Omega} \frac{(u-v)^2}{2f} \, d\vec{x} + \int_{\Omega} (v-u) \frac{f-u}{f} \, d\vec{x}. \end{aligned}$$

Now for any $t \in \mathbb{R}$, we get

$$\lambda(J(u+t(v-u)) - J(u)) - t \int_{\Omega} (v-u) \frac{f-u}{f} \, d\vec{x} \geq -\frac{t^2}{2} \int_{\Omega} \frac{(v-u)^2}{f} \, d\vec{x}.$$

In the above inequality, the left-hand side is a convex function of $t \in \mathbb{R}$ and the right-hand side is a concave parabola (as a function of t). The maximum point of the parabola is at $t = 0$, and it meets the convex function at this point. Thus, we can easily conclude that the convex function on the left-hand side will be larger than the maximum of the parabola, which is zero here, at every point. Hence,

$$\lambda(J(u+t(v-u)) - J(u)) - t \int_{\Omega} (v-u) \frac{f-u}{f} \, d\vec{x} \geq 0.$$

Since this is true for any $t \in \mathbb{R}$, considering $f \approx u$ and $t = 1$ gives us, for all $v \in L^2(\Omega)$,

$$J(v) \geq J(u) + \int_{\Omega} (v-u) \frac{f-u}{\lambda u} \, d\vec{x}.$$

Using the definition of subdifferential [Chambolle 2004],

$$\frac{f-u}{\lambda u} \in \partial J(u).$$

Conversely, if this is true, then we see that (18) holds.

Thus u is a minimizer of (16), and (18) gives the required Euler–Lagrange equation. \square

The MCSD algorithm. From the Euler–Lagrange equation (18) and the Legendre–Fenchel identity property,

$$u \in \partial J^* \left(\frac{f-u}{\lambda u} \right).$$

Setting $w = (f - u)/(\lambda u)$, we have

$$\begin{aligned} u \in \partial J^*(w) &\Rightarrow \frac{f}{\lambda u} \in \frac{f - u}{\lambda u} + \frac{\partial J^*(w)}{\lambda u} \\ &\Rightarrow 0 \in w - \frac{f}{\lambda u} + \frac{\partial J^*(w)}{\lambda u} \end{aligned} \tag{19}$$

$$\Rightarrow 0 \in uw - \frac{f}{\lambda} + \frac{\partial J^*(w)}{\lambda}. \tag{20}$$

If the minimizing functional, where w is the minimizer, is given by

$$\frac{\|\sqrt{u}w - f/(\lambda\sqrt{u})\|^2}{2} + \frac{1}{\lambda}J^*(w), \tag{21}$$

then similar analysis as in the proof of Theorem 4.1 yields that the corresponding Euler–Lagrange equation is (20).

Since $w \in K$, where $K = \{p \in Y, \|p_{i,j}\| \leq 1 \forall i, j\}$, and $J^* = H$, where H is defined as

$$H(w) = \begin{cases} 0 & \text{if } w \in K, \\ +\infty & \text{if } w \notin K, \end{cases} \tag{22}$$

we get $J^*(w) = 0$. Therefore, finding a minimizer w for (21) is equivalent to solving the problem

$$\min_p \left\{ \left\| \sqrt{u} \operatorname{div} p - \frac{f}{\lambda\sqrt{u}} \right\|^2 : p \in Y, |p_{i,j}| \leq 1 \forall i, j = 1, \dots, N \right\}. \tag{23}$$

We need to find p for (23) and then recover u by

$$u = f - \lambda u \operatorname{div} p.$$

Now, to minimize (23), we consider

$$-\left[\nabla \left(\sqrt{u} \operatorname{div} p - \frac{f}{\lambda\sqrt{u}} \right) \right]_{i,j} + \alpha_{ij} p_{ij} = 0,$$

where $\alpha_{ij} \geq 0$ is a Lagrange multiplier. One can verify that $\alpha_{ij} > 0$ and $|p_{ij}| = 1$, or $|p_{ij}| < 1$ and $\alpha_{ij} = 0$. In any case,

$$\alpha_{ij} = \left| \left(\nabla \left(\sqrt{u} \operatorname{div} p^n - \frac{f}{\lambda\sqrt{u}} \right) \right)_{i,j} \right|.$$

Applying gradient descent, we can obtain the solution iteratively by the semi-implicit algorithm

$$p_{i,j}^{n+1} = p_{i,j}^n + \tau \left(\left(\nabla \left(\sqrt{u} \operatorname{div} p^n - \frac{f}{\lambda\sqrt{u}} \right) \right)_{i,j} - \left| \left(\nabla \left(\sqrt{u} \operatorname{div} p^n - \frac{f}{\lambda\sqrt{u}} \right) \right)_{i,j} \right| p_{i,j}^{n+1} \right) \tag{24}$$

for $n \geq 0$, $p^0 = 0$, and for an iterative time-step size $\tau > 0$. Then (24) gives

$$p_{i,j}^{n+1} = \frac{p_{i,j}^n + \tau (\nabla(\sqrt{u} \operatorname{div} p^n - f/(\lambda\sqrt{u}))_{i,j})}{1 + \tau |(\nabla(\sqrt{u} \operatorname{div} p^n - f/(\lambda\sqrt{u}))_{i,j})|}.$$

For the TV model, Chambolle proposed updating λ for denoising purposes using (13):

$$\lambda^{n+1} = \frac{\|f - u_c\|}{\|\operatorname{div} p^{n+1}\|}.$$

Here, u_c is the noise-free clear image. But this can only be obtained for synthetic images. For speckle noise images, we change (13) to

$$\lambda^{n+1} = \frac{\|f - f_s\|}{\|u^n \operatorname{div} p^{n+1}\|},$$

where f_s is the smoother version of the original or given image. For any (i, j) , we obtain $f_s(i, j)$ by considering the average of the four surrounding pixels. Hence, the iterative algorithm for MCSD is given for $n \geq 0$ as

$$p_{i,j}^{n+1} = \frac{p_{i,j}^n + \tau (\nabla(\sqrt{u^n} \operatorname{div} p^n - f/(\lambda\sqrt{u^n}))_{i,j})}{1 + \tau |(\nabla(\sqrt{u^n} \operatorname{div} p^n - f/(\lambda\sqrt{u^n}))_{i,j})|}, \quad (25)$$

$$\lambda^{n+1} = \frac{\|f - f_s\|}{\|u^n \operatorname{div} p^{n+1}\|}, \quad (26)$$

$$u^{n+1} = f - \lambda^{n+1} u^n \operatorname{div} p^{n+1}. \quad (27)$$

Note that the problem satisfies the zero Neumann boundary condition

$$\frac{\partial u}{\partial n} = 0 \text{ on } \partial\Omega. \quad (28)$$

5. Numerical procedure: the MCSD model

We now describe the numerical procedure used for the MCSD model. In [Chambolle 2004], the discrete gradient and divergence were defined using forward and backward differences respectively. Also, Chambolle used separate definitions for boundary points. In this model, we use central differences to define both the gradient and divergence, as we have seen that this gives more accurate results. Also, we introduce 1-pixel-wide ghost grids on each side to avoid separate definitions for boundary points and also to satisfy the zero Neumann boundary condition (28). The ghost grid values are obtained by

$$\begin{aligned} u(0, :) &= u(2, :), & u(N+1, :) &= u(N-1, :), \\ u(:, 0) &= u(:, 2), & u(:, N+1) &= u(:, N-1). \end{aligned}$$

For $u \in X = \mathbb{R}^{N \times N}$, we have

$$(\nabla u)_{i,j} = ((\nabla u)_{i,j}^1, (\nabla u)_{i,j}^2),$$

where

$$(\nabla u)_{i,j}^1 = \frac{1}{2}(u_{i+1,j} - u_{i-1,j}), \quad (\nabla u)_{i,j}^2 = \frac{1}{2}(u_{i,j+1} - u_{i,j-1})$$

for $1 \leq i, j \leq N$. For any vector $p = (p^1, p^2) \in Y = X \times X$, we define

$$(\operatorname{div} p)_{i,j} = \frac{1}{2}(p_{i+1,j}^1 - p_{i-1,j}^1) + \frac{1}{2}(p_{i,j+1}^2 - p_{i,j-1}^2) \tag{29}$$

for $1 \leq i, j \leq N$. Applying these definitions, we solve the iterative algorithm provided in (25)–(27).

Convergence analysis. For purposes of convergence analysis of the MCSD algorithm developed on page 332, we slightly modify the algorithm by using (19) for the Euler–Lagrange equation instead of (20). We do not iterate u in the algorithm by replacing it by f_s , where f_s is the smoother version of the original or given image. Thus the modified Euler–Lagrange equation is given by

$$0 \in w - \frac{f}{\lambda f_s} + \frac{\partial J^*(w)}{\lambda f_s}.$$

The corresponding minimizing functional is therefore

$$\frac{1}{2} \left\| w - \frac{f}{\lambda f_s} \right\|^2 + \frac{1}{\lambda f_s} J^*(w), \tag{30}$$

where $w = (f - u)/(\lambda u)$ is the minimizer. Thus we obtain the semi-implicit algorithm

$$p_{i,j}^{n+1} = p_{i,j}^n + \tau \left(\left(\nabla \left(\operatorname{div} p^n - \frac{f}{\lambda f_s} \right) \right)_{i,j} - \left| \left(\nabla \left(\operatorname{div} p^n - \frac{f}{\lambda f_s} \right) \right)_{i,j} \right| p_{i,j}^{n+1} \right) \tag{31}$$

and that

$$p_{i,j}^{n+1} = \frac{p_{i,j}^n + \tau (\nabla (\operatorname{div} p^n - f/(\lambda f_s)))_{i,j}}{1 + \tau \left| (\nabla (\operatorname{div} p^n - f/(\lambda f_s)))_{i,j} \right|} \tag{32}$$

for $n \geq 0$, $p^0 = 0$, and for an iterative time-step size $\tau > 0$. Then the following convergence theorem holds.

Theorem 5.1. *If $\tau \leq \frac{1}{2}$, then $f - \lambda f_s \operatorname{div} p^n$, where p^n is obtained by (32), converges to u as $n \rightarrow \infty$.*

Proof. Fix $n \geq 0$ and let $\eta = (p^{n+1} - p^n)/\tau$. Since

$$\operatorname{div} p^{n+1} - \frac{f}{\lambda f_s} = \operatorname{div} p^n - \frac{f}{\lambda f_s} + \tau \operatorname{div} \eta,$$

we derive the following:

$$\begin{aligned} \left\| \operatorname{div} p^{n+1} - \frac{f}{\lambda f_s} \right\|^2 &= \left\| \operatorname{div} p^n - \frac{f}{\lambda f_s} + \tau \operatorname{div} \eta \right\|^2 \\ &= \left\| \operatorname{div} p^n - \frac{f}{\lambda f_s} \right\|^2 + 2\tau \left\langle \operatorname{div} \eta, \operatorname{div} p^n - \frac{f}{\lambda f_s} \right\rangle + \tau^2 \|\operatorname{div} \eta\|^2. \end{aligned}$$

Let κ be the norm of $\operatorname{div} : Y \rightarrow X$. That is, $\kappa = \sup_{\|p\|_Y \leq 1} \|\operatorname{div} p\|_X$. Then, we get $\|\operatorname{div} \eta\|^2 \leq \kappa^2 \|\eta\|_Y^2$. Using this and also by the property $\langle \operatorname{div} p, u \rangle = -\langle p, \nabla u \rangle$, we obtain the inequality

$$\left\| \operatorname{div} p^{n+1} - \frac{f}{\lambda f_s} \right\|^2 \leq \left\| \operatorname{div} p^n - \frac{f}{\lambda f_s} \right\|^2 - \tau \left(2 \left\langle \eta, \nabla \left(\operatorname{div} p^n - \frac{f}{\lambda f_s} \right) \right\rangle - \kappa^2 \tau \|\eta\|_Y^2 \right).$$

Note that

$$2 \left\langle \eta, \nabla \left(\operatorname{div} p^n - \frac{f}{\lambda f_s} \right) \right\rangle - \kappa^2 \tau \|\eta\|_Y^2 = \sum_{i,j=1}^N 2\eta_{i,j} \left(\nabla \left(\operatorname{div} p^n - \frac{f}{\lambda f_s} \right) \right)_{i,j} - \kappa^2 \tau \eta_{i,j}^2$$

and by (31),

$$\eta_{i,j} = \left(\nabla \left(\operatorname{div} p^n - \frac{f}{\lambda f_s} \right) \right)_{i,j} - \rho_{i,j},$$

where $\rho_{i,j} = |(\nabla(\operatorname{div} p^n - f/(\lambda f_s)))_{i,j}| p_{i,j}^{n+1}$. Let $a_{i,j} := (\nabla(\operatorname{div} p^n - f/(\lambda f_s)))_{i,j}$. Then

$$\begin{aligned} 2\eta_{i,j} \left(\nabla \left(\operatorname{div} p^n - \frac{f}{\lambda f_s} \right) \right)_{i,j} - \kappa^2 \tau \eta_{i,j}^2 &= 2\eta_{i,j} a_{i,j} - 2\eta_{i,j} \rho_{i,j} + 2\eta_{i,j} \rho_{i,j} - \kappa^2 \tau \eta_{i,j}^2 \\ &= 2\eta_{i,j}^2 + 2\eta_{i,j} \rho_{i,j} - \kappa^2 \tau \eta_{i,j}^2 \\ &= (1 - \kappa^2 \tau) \eta_{i,j}^2 + \eta_{i,j}^2 + 2\eta_{i,j} \rho_{i,j} \\ &= (1 - \kappa^2 \tau) \eta_{i,j}^2 + a_{i,j}^2 - \rho_{i,j}^2. \end{aligned}$$

Since $p^0 = 0$, we can easily prove by induction that $|p_{i,j}^n| \leq 1 \forall i, j$ for all $n \geq 0$. This implies $\rho_{i,j} \leq |a_{i,j}|$. Therefore, if $\tau \leq 1/\kappa^2$, then

$$2 \left\langle \eta, \nabla \left(\operatorname{div} p^n - \frac{f}{\lambda f_s} \right) \right\rangle - \kappa^2 \tau \|\eta\|_Y^2 \geq 0,$$

which implies $\|\operatorname{div} p^n - f/(\lambda f_s)\|^2$ is decreasing with n . Hence, there exists a limit of $\|\operatorname{div} p^n - f/(\lambda f_s)\|^2$ as $n \rightarrow \infty$ and we can conclude $f - \lambda f_s \operatorname{div} p^n$ converges to a solution of the simplified version of the MCSD minimizing functional (21).

Now we need to show that $\tau \leq 1/\kappa^2$ if $\tau \leq \frac{1}{2}$. By (29),

$$\begin{aligned} \|\operatorname{div} p\|^2 &= \sum_{1 \leq i, j \leq N} \left(\frac{1}{2}(p_{i+1, j}^1 - p_{i-1, j}^1) + \frac{1}{2}(p_{i, j+1}^2 - p_{i, j-1}^2) \right)^2 \\ &\leq \sum_{1 \leq i, j \leq N} (p_{i+1, j}^1)^2 + (p_{i-1, j}^1)^2 + (p_{i, j+1}^2)^2 + (p_{i, j-1}^2)^2 \\ &\leq 2\|p\|_Y^2. \end{aligned}$$

This proves that $\kappa^2 \leq 2$. Since we assumed $\tau \leq \frac{1}{2}$, we finally get $\tau \leq 1/\kappa^2$. \square

6. Numerical results for the MCSD model

The resulting images shown here are obtained using C++ programs which were compiled and run on Linux. Other than comparing the visual results, we also use peak-signal-to-noise ratio (PSNR) to measure image quality. The definition of PSNR is given as follows:

Definition 6.1 (PSNR). Let g be a noise-free clean image and u be the restored image obtained by denoising a noisy version of g . The PSNR is measured by

$$\text{PSNR} = 10 \log_{10} \left(\frac{\sum_{ij} 255^2}{\sum_{ij} (g_{ij} - u_{ij})^2} \right).$$

Note that if the denoised image is very close to the clean image, the denominator will be very small, thus providing a higher PSNR for a cleaner image. Also, the PSNR can be obtained for images with synthetically added noise only. Images with natural noise cannot have PSNR values since g is not available.

First we show the results for a Gaussian noise image (synthetic Lena image). From Figure 1, we can compare the results obtained from different denoising models. The MCSD model does not produce a nice result, as it is meant for speckle denoising. The Chambolle with central difference scheme provides the best denoised image. This fact is also supported by the PSNR values in Table 1, where the central difference Chambolle model has the highest PSNR and MCSD has the lowest, but the MCSD model is still faster than the TV model (6).

Next we have the results together with the residuals for the MCSD (25)–(27) and the Krissian et al. (15) models for speckle noise images. In Figure 2 (the speckle Lena image), we see that MCSD has comparable results to the Krissian et al. model. The residuals in Figure 3 show that MCSD has picked up more noise and less detail than the Krissian et al. model. From Table 2, we see that MCSD has a higher PSNR value than Krissian et al. and it also takes less time.

Figure 4 shows the results for an ultrasound image (liver image). Here also we can see MCSD performs better than Krissian et al. The residual image (Figure 5)



noise-free image



Gaussian noise image



TV



Chambolle



Chambolle (cdm)



MCSD

Figure 1. Results of TV, Chambolle and MCSD models (Gaussian Lena image).

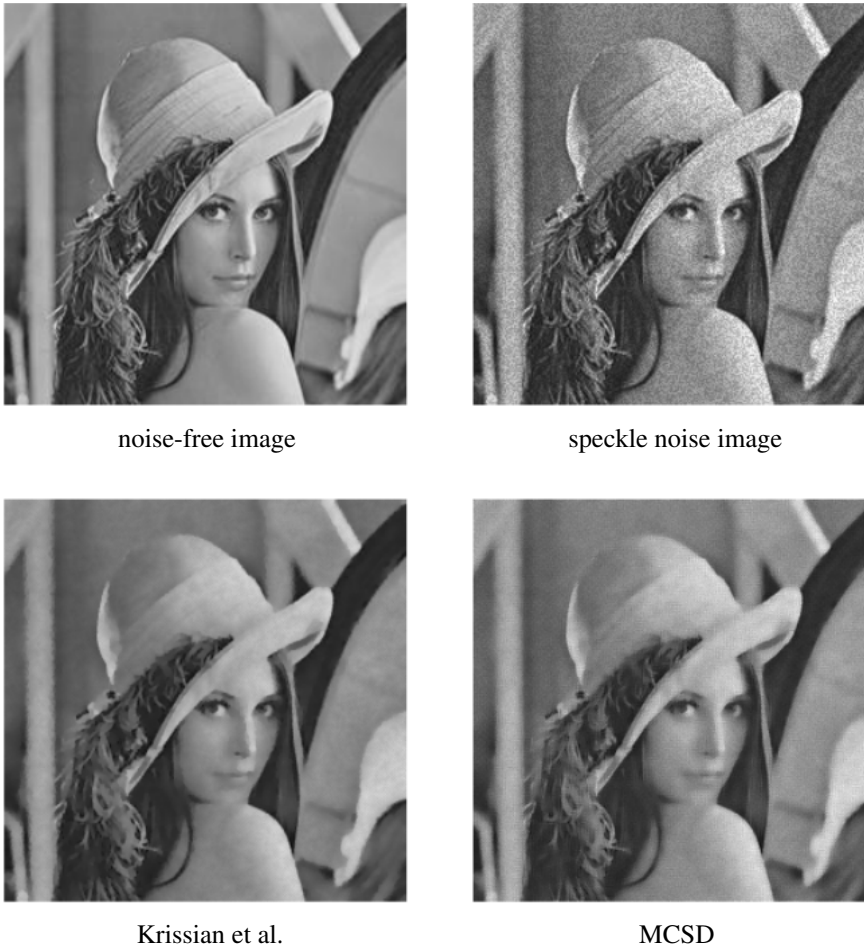


Figure 2. Results of Krissian et al. and MCSD models (speckle Lena image).

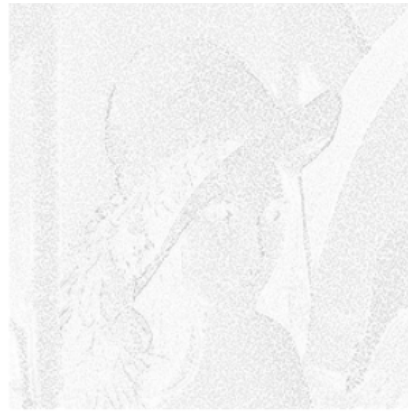
TV		Chambolle		Chambolle (cdm)		MCSD	
time	PSNR	time	PSNR	time	PSNR	time	PSNR
13.71	27.78	4.54	27.71	4.64	28.86	8.98	26.06

Table 1. Model comparison for the Lena image with Gaussian noise and PSNR = 24.30, where time is measured in seconds.

shows MCSD picked up more noise but preserved more edges compared to the Krissian et al. model. We are unable to compare PSNR values here since they are images having natural noise. But Table 2 does show us that MCSD is faster than the Krissian et al. model.



Krissian et al. absolute residual



MCSD absolute residual



Krissian et al. noise residual



MCSD noise residual

Figure 3. Residuals of Krissian et al. and MCSD models (speckle Lena image).

images	Krissian et al.		MCSD	
	time	PSNR	time	PSNR
Lena (PSNR= 25.70)	2.83	27.02	0.14	28.52
liver	2.35	—	0.13	—

Table 2. Model comparison for speckle noise, where time is measured in seconds.



noisy image



Krissian et al.

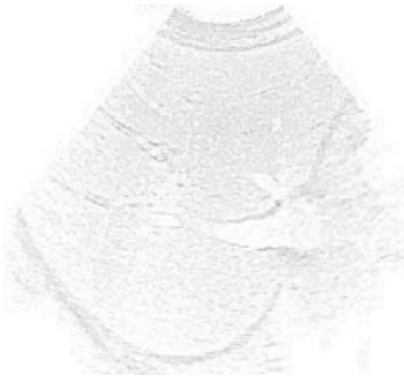


MCSD

Figure 4. Results of Krissian et al. and MCSD models (ultrasound liver image).

7. Conclusion

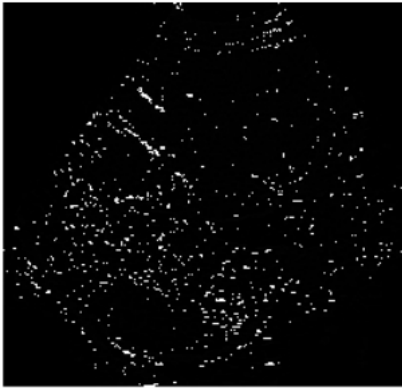
In this paper, we introduced our new TV based denoising model for speckle noise images. The new model provides a speckle noise version for the Chambolle algorithm, which was originally designed for faster solution of the ROF model. The results show a significant amount of improvement compared to the conventional TV based speckle denoising model. Based on a dual formation, the solution is updated directly from the dual space. The new method is therefore much more efficient than the method by Krissian et al. It is also numerically shown that the new method is more accurate than the Krissian et al. method. Under certain conditions on the time-step size, it is proved that the solution from the new algorithm converges to the minimizer of the new speckle denoising model.



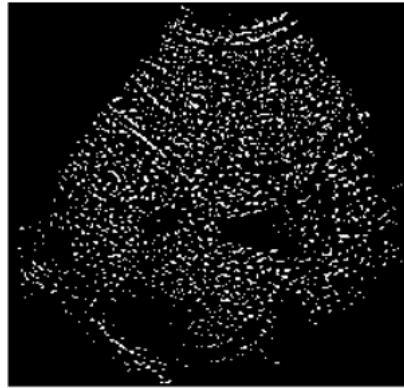
Krissian et al. absolute residual



MCSD absolute residual



Krissian et al. noise residual



MCSD noise residual

Figure 5. Residuals of Krissian et al. and MCSD models (ultrasound liver image).

Acknowledgements

This work is supported in part by National Science Foundation grant DMS-0852032 and by the Saginaw Valley State University Faculty Research Grant for Arundhati Bagchi Misra.

References

- [Alvarez et al. 1992] L. Alvarez, P.-L. Lions, and J.-M. Morel, “Image selective smoothing and edge detection by nonlinear diffusion, II”, *SIAM J. Numer. Anal.* **29**:3 (1992), 845–866. MR Zbl
- [Carter 2001] J. L. Carter, *Dual methods for total variation-based image restoration*, Ph.D thesis, University of California, Los Angeles, 2001, <http://search.proquest.com/docview/304689850>. MR

- [Catté et al. 1992] F. Catté, P.-L. Lions, J.-M. Morel, and T. Coll, “Image selective smoothing and edge detection by nonlinear diffusion”, *SIAM J. Numer. Anal.* **29**:1 (1992), 182–193. MR Zbl
- [Chambolle 2004] A. Chambolle, “An algorithm for total variation minimization and applications”, *J. Math. Imaging Vision* **20**:1–2 (2004), 89–97. MR
- [Chambolle and Pock 2011] A. Chambolle and T. Pock, “A first-order primal-dual algorithm for convex problems with applications to imaging”, *J. Math. Imaging Vision* **40**:1 (2011), 120–145. MR Zbl
- [Chambolle et al. 2010] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock, “An introduction to total variation for image analysis”, pp. 263–340 in *Theoretical foundations and numerical methods for sparse recovery*, Radon Ser. Comput. Appl. Math. **9**, Walter de Gruyter, Berlin, 2010. MR Zbl
- [Chan and Vese 1997] T. Chan and L. Vese, “Variational image restoration and segmentation models and approximations”, CAM report 97-47, University of California, Los Angeles, 1997, <ftp://ftp.math.ucla.edu/pub/.../cam97-47.ps.gz>.
- [Chan et al. 1999] T. F. Chan, G. H. Golub, and P. Mulet, “A nonlinear primal-dual method for total variation-based image restoration”, *SIAM J. Sci. Comput.* **20**:6 (1999), 1964–1977. MR Zbl
- [Chan et al. 2014] R. Chan, H. Yang, and T. Zeng, “A two-stage image segmentation method for blurry images with Poisson or multiplicative gamma noise”, *SIAM J. Imaging Sci.* **7**:1 (2014), 98–127. MR Zbl
- [Dong and Zeng 2013] Y. Dong and T. Zeng, “A convex variational model for restoring blurred images with multiplicative noise”, *SIAM J. Imaging Sci.* **6**:3 (2013), 1598–1625. MR Zbl
- [Huang et al. 2012] Y.-M. Huang, L. Moisan, M. K. Ng, and T. Zeng, “Multiplicative noise removal via a learned dictionary”, *IEEE Trans. Image Process.* **21**:11 (2012), 4534–4543. MR
- [Huang et al. 2013a] Y. Huang, M. Ng, and T. Zeng, “The convex relaxation method on deconvolution model with multiplicative noise”, *Commun. Comput. Phys.* **13**:4 (2013), 1066–1092. MR
- [Huang et al. 2013b] Y.-M. Huang, D.-Y. Lu, and T. Zeng, “Two-step approach for the restoration of images corrupted by multiplicative noise”, *SIAM J. Sci. Comput.* **35**:6 (2013), A2856–A2873. MR Zbl
- [Jin and Yang 2011] Z. Jin and X. Yang, “A variational model to remove the multiplicative noise in ultrasound images”, *J. Math. Imaging Vision* **39**:1 (2011), 62–74. MR Zbl
- [Joo and Kim 2003a] K. Joo and S. Kim, “PDE-based image restoration, I: Anti-staircasing and anti-diffusion”, research report, University of Kentucky, 2003.
- [Joo and Kim 2003b] K. Joo and S. Kim, “PDE-based image restoration, II: Numerical schemes and color image denoising”, research report, University of Kentucky, 2003.
- [Kim 2004] S. Kim, “Loss and recovery of fine structures in pde-based image denoising”, September 6–9 2004, <https://www.ceremade.dauphine.fr/~cohen/mia2004/>. Talk at the fifth conference on Mathematics and Image Analysis, Paris.
- [Kim and Lim 2007] S. Kim and H. Lim, “A non-convex diffusion model for simultaneous image denoising and edge enhancement”, pp. 175–192 in *Proceedings of the Sixth Mississippi State-UBA Conference on Differential Equations and Computational Simulations*, edited by J. Graef et al., Electron. J. Differ. Equ. Conf. **15**, Southwest Texas State Univ., San Marcos, TX, 2007. MR Zbl
- [Kornprobst et al. 1997] P. Kornprobst, R. Deriche, and G. Aubert, “Nonlinear operators in image restoration”, pp. 325–331 in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (San Juan, 1997), 1997.

- [Krissian et al. 2005] K. Krissian, R. Kikinis, C.-F. Westin, and K. Vosburgh, “Speckle-constrained filtering of ultrasound images”, pp. 547–552 in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Washington D.C., 2005), vol. 2, 2005.
- [Marquina and Osher 2000] A. Marquina and S. Osher, “Explicit algorithms for a new time dependent model based on level set motion for nonlinear deblurring and noise removal”, *SIAM J. Sci. Comput.* **22**:2 (2000), 387–405. MR Zbl
- [Perona and Malik 1990] P. Perona and J. Malik, “Scale-space and edge detection using anisotropic diffusion”, *IEEE Trans. Pattern Anal. Mach. Intell.* **12**:7 (1990), 629–639.
- [Rudin et al. 1992] L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms”, *Phys. D* **60**:1–4 (1992), 259–268. MR Zbl
- [Vese and Chan 1997] L. Vese and T. Chan, “Reduced non-convex functional approximations for image restoration & segmentation”, CAM report 97-56, University of California, Los Angeles, 1997, <ftp://ftp.math.ucla.edu/pub/.../cam97-56.ps.gz>.
- [Wen et al. 2016] Y. Wen, R. H. Chan, and T. Zeng, “Primal-dual algorithms for total variation based image restoration under Poisson noise”, *Sci. China Math.* **59**:1 (2016), 141–160. MR Zbl

Received: 2015-12-19 Revised: 2016-02-26 Accepted: 2016-03-19

abmisra@svsu.edu *Department of Mathematical Sciences, Saginaw Valley
State University, University Center, MI 48710, United States*

elockhart@math.arizona.edu *Program in Applied Mathematics, The University of Arizona,
Tuscon, AZ 85721, United States*

hlim@math.msstate.edu *Department of Mathematics and Statistics, Mississippi
State University, Mississippi State, MS 39762, United States*

A new look at Apollonian circle packings

Isabel Corona,Carolynn Johnson, Lon Mitchell and Dylan O’Connell

(Communicated by Scott T. Chapman)

We define an abstract Apollonian supergasket using the solution set of a certain Diophantine equation, showing that the solutions are in bijective correspondence with the circles of any concrete supergasket. Properties of the solution set translate directly to geometric and algebraic properties of Apollonian gaskets, facilitating their study. In particular, curvatures of individual circles are explored and geometric relationships among multiple circles are given simple algebraic expressions. All results can be applied to a concrete gasket using the curvature-center coordinates of its four defining circles. These techniques can also be applied to other types of circle packings and higher-dimensional analogs.

An Apollonian gasket is a type of circle packing in the plane generated recursively starting from a set of four mutually tangent circles. The curvatures of any four such circles are related by an equation discovered by Descartes, and every circle in a gasket generated by four circles with integer curvatures will have integer curvature. While these gaskets have been fascinating to mathematicians for some time—the use of group theory in their study was initiated by Keith Hirst [1967] and they even inspired a poem¹—it was only relatively recently that Jeffrey Lagarias, Colin Mallows, and Allan Wilks [Lagarias et al. 2002] gave an algebraic characterization of Descartes configurations. One question in particular has inspired much work but resisted a complete answer: given the four original integer curvatures, which other curvatures can or will occur, and how frequently? Peter Sarnak [2011], Elena Fuchs [2013], and Hee Oh [2014] have recent surveys on this topic, which has seen significant progress in the past five years [Bourgain 2012; Bourgain and Kontorovich 2014; Bourgain and Fuchs 2011; Fuchs and Sanden 2011].

In this paper, inspired by recent work of Sam Northshield [2015], we provide a four-dimensional label to each circle that does not depend on the location of the circle but refers instead to its geometric relationship to the original four circles. Since we consider only the process of generating the gasket, the labels provide an abstract version of an Apollonian circle packing that can represent any concrete

MSC2010: primary 52C26; secondary 11D09.

Keywords: Apollonian circle packing, Apollonian gasket, Apollonian supergasket.

¹*The Kiss Precise* by Frederick Soddy, 1936.

packing once an initial set of four circles is specified. These labels can be used to determine location and radius, find whether given circles in a gasket are tangent or not, perform operations such as inversion, and obtain curvature results. This technique is equally applicable to any packing generated in a similar fashion, such as the generalizations of Apollonian packings of Gerhard Guettler and Colin Mallows [2010] or packings in higher-dimensional Euclidean, spherical, or hyperbolic spaces [Lagarias et al. 2002].

1. Descartes configurations

Descartes configurations are the basic building blocks of Apollonian circle packings. We begin by providing a brief introduction; for more detail, see the paper by Lagarias, Mallows, and Wilks [Lagarias et al. 2002] or any of the surveys mentioned above.

An oriented circle in the plane consists of a circle and an orientation, thought of as a unit normal vector, of “inward” or “outward” that specifies its interior. The curvature of a circle is the inverse of its radius; the oriented curvature of an oriented circle is the curvature if the circle has an inward-pointing normal vector and the negative of the curvature otherwise. Two circles are tangent if they intersect in a single point. Lines are considered to be circles of curvature zero, and two lines that are not the same are considered to be tangent at infinity. In what follows, by a circle we will mean either an oriented circle or oriented line, tangent will mean externally tangent, and by the curvature of a circle, we will mean the oriented curvature.

A *Descartes configuration* (hereafter, configuration) consists of four circles in the plane that are pairwise externally tangent and such that no three share a point of tangency. There are four basic types of configurations, shown in Figure 1. Descartes discovered that the oriented curvatures κ_i of four oriented circles in a configuration satisfy

$$2(\kappa_1^2 + \kappa_2^2 + \kappa_3^2 + \kappa_4^2) = (\kappa_1 + \kappa_2 + \kappa_3 + \kappa_4)^2, \quad (1)$$

which we will call the *Descartes condition*.²

The Descartes condition is not enough to characterize configurations, but a characterization exists using additional information [Graham et al. 2005; Lagarias et al. 2002], and the geometry of inversion over a circle plays an important part. For a line, inversion over the line is simply reflection. For a circle C with center O and radius r , inversion over C is the Möbius transformation I_C that maps a point P to the point Q on the ray from O through P such that $r^2 = |OP||OQ|$. Each inversion is anticonformal in that it preserves magnitudes of angles but reverses their directions; further, inversion over a circle or line maps oriented circles and lines to oriented circles and lines.

²Descartes considered configurations without lines, but with our definitions, (1) is true for any type of configuration [Lagarias et al. 2002].

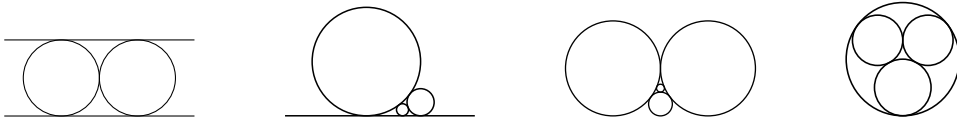


Figure 1. Descartes configurations.

Each circle that is not a line is uniquely identified by its center and curvature, since the curvature provides both radius and orientation. To uniquely identify all circles, Lagarias, Mallows, and Wilks devised *curvature-center coordinates*, which for any circle are of the form k', k, x, y , where k is the curvature and k' is the curvature of the inversion of the circle over the unit circle; if the curvature k is nonzero, then $x = kc_x$ and $y = kc_y$, where (c_x, c_y) is the center of the circle; if the curvature k is zero, then x and y are the corresponding components of the unit normal vector. For example, the curvature-center coordinates of the unit circle with the origin in its interior are $-1, 1, 0, 0$ and the curvature-center coordinates of the line $y = 1$ with the origin in its interior are $2, 0, 0, -1$.

Here is the characterization of configurations: let C_1, \dots, C_4 be circles, let $M = M(C_1, \dots, C_4)$ be the *curvature-center matrix* of the circles C_1, \dots, C_4 , where each row consists of the curvature-center coordinates of the corresponding circle, and let

$$Q = \begin{bmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 \end{bmatrix}.$$

(Our Q matrix is twice the Q of Lagarias et al. [2002] for notational convenience.)

Theorem 1 (augmented Euclidean Descartes theorem [Lagarias et al. 2002; Graham et al. 2005]). *Circles C_1, \dots, C_4 form a configuration if and only if*

$$M^T Q M = \begin{bmatrix} 0 & -8 & 0 & 0 \\ -8 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix} =: W. \tag{2}$$

Note that the matrix Q is related to the Descartes condition in that if $\vec{x} = (x_1 \ x_2 \ x_3 \ x_4)^T$ is a column vector then

$$\langle \vec{x}, \vec{x} \rangle_Q := \vec{x}^T Q \vec{x} = 2(x_1^2 + x_2^2 + x_3^2 + x_4^2) - (x_1 + x_2 + x_3 + x_4)^2.$$

Indeed, the first two diagonal entries of W correspond to the Descartes condition.

Given any three mutually tangent circles $C_1, C_2,$ and C_3 that do not share a point of tangency, there are exactly two other circles that each form a configuration with the original three [Sarnak 2011]. The operation that takes a configuration

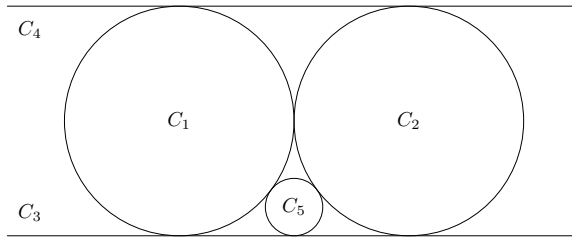


Figure 2. An example of reflection.

C_1, C_2, C_3, C_4 to the configuration C_1, C_2, C_3, C_5 is defined to be the *reflection* (of C_4 over $C_1, C_2,$ and C_3) [Graham et al. 2005] (and when the context allows we will speak of replacing C_4 with C_5 in this fashion). In Figure 2, for example, C_5 is the reflection of C_4 over $C_1, C_2,$ and C_3 (and C_4 is the reflection of C_5 over $C_1, C_2,$ and C_3), and hence we can speak of replacing C_4 in the configuration C_1, C_2, C_3, C_4 with C_5 to obtain the configuration C_1, C_2, C_3, C_5 .

Since inversion over a circle preserves tangency, inverting three circles of a configuration over the fourth will also result in another Descartes configuration. For example, in Figure 3, the three smallest circles invert over circle C_1 to the three largest circles.

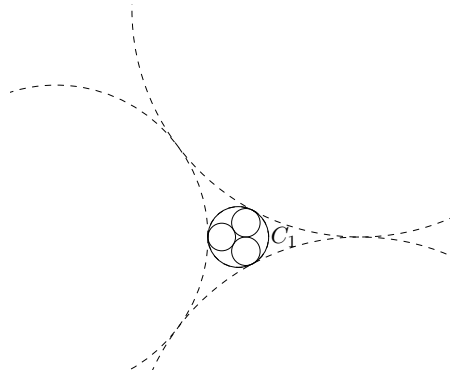


Figure 3. An example of inversion.

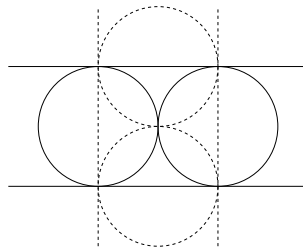


Figure 4. A configuration (solid lines) and its dual (dashed lines).

Finally, each configuration C_1, \dots, C_4 also has a dual configuration C'_1, \dots, C'_4 such that each C'_i does not intersect C_i and goes through the three points of tangency of the other three C_j with $j \neq i$. For example, in Figure 4, a configuration (solid lines) is superimposed with its dual (dashed lines).

2. Apollonian gaskets

Apollonian Gaskets can be defined geometrically and algebraically. In this section, we will review the geometric construction.

Given three mutually tangent circles, there are exactly two other circles that form a configuration with the original three. Thus, starting with a configuration of four circles, any three of the four define a new configuration not including the other circle. Repeatedly creating new configurations in this fashion, a circle packing (a collection of circles with mutually disjoint interiors) is created, called an Apollonian circle packing or Apollonian gasket; see Figure 5.

If $\kappa_1, \dots, \kappa_5$ are the curvatures of five circles C_1, \dots, C_5 such that C_1, C_2, C_3, C_4 and C_1, C_2, C_3, C_5 are configurations, the Descartes condition implies

$$\kappa_5 = 2\kappa_1 + 2\kappa_2 + 2\kappa_3 - \kappa_4. \tag{3}$$

Thus, in an Apollonian gasket, because each circle belongs to a configuration that can be obtained from the original one by repeated replacement operations, if the original curvatures are integers then the curvatures of all the circles in the gasket will also be integers.

The gasket with starting curvatures 0, 0, 2, and 2 contains another set of well-known circles called the Ford circles, shown in Figure 6, which can be defined as follows. For $r > 0$ and arbitrary real a , let $C(a, r)$ be the circle with radius r above and tangent to the x -axis at $x = a$. For relatively prime integers c and d with $d \neq 0$, let $C_{c,d} = C(c/d, 1/(2d^2))$; the set of all such $C_{c,d}$ are the Ford circles. These circles have a number of interesting properties. To see they are part of the (2, 2, 0, 0)-gasket (which we will call the *Ford gasket*) invokes one of these properties: if $C_{a,b}$ and $C_{c,d}$ are mutually tangent, then $C_{a+c,b+d}$ forms a

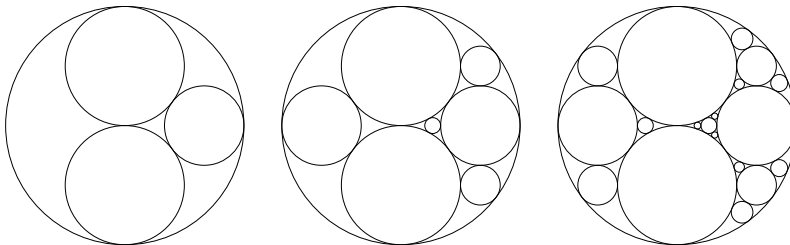


Figure 5. An Apollonian gasket.

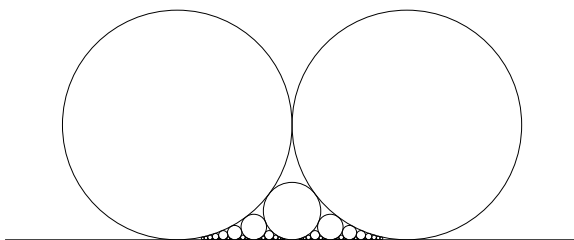


Figure 6. Ford circles.

Descartes configuration with $C_{a,b}$, $C_{c,d}$, and the x -axis. The claim then follows from $C_{0,1}$, $C_{1,1}$, and the x -axis being part of the original four gasket circles.

Sam Northshield [2015] recently discovered a new characterization and labeling for the Ford circles. For integers s and t with $s + t > 0$, define

$$\langle s, t \rangle = C\left(\frac{s}{s+t}, \frac{1}{(s+t)^2}\right).$$

Then the set of Ford circles is exactly the set of those $\langle s, t \rangle$ with integer s and t that satisfy two conditions: $s + t > 0$ and there exists an integer u such that $\gcd(s, t, u) = 1$ and $s^2 + t^2 + u^2 = (s + t + u)^2$. This characterization also allowed Northshield to study natural generalizations of the Ford circles in higher dimensions.

3. The Apollonian group

Each geometric operation described above has a matrix counterpart. For example, consider two configurations C_1, C_2, C_3, C_4 and C_1, C_2, C_3, C_5 , let

$$S_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 2 & 2 & 2 & -1 \end{bmatrix}$$

and let $M = M(C_1, C_2, C_3, C_4)$. We claim that $S_4M = M(C_1, C_2, C_3, C_5)$. Since $S_4^T Q S_4 = Q$, we have $(S_4M)^T Q S_4M = M^T Q M = W$, so that S_4M is also a configuration. Since S_4 does not change C_1, C_2 , or C_3 , it follows that S_4M must be the unique configuration obtained by reflection of C_4 . This provides an alternate way of defining an Apollonian gasket.

The Apollonian group \mathcal{A} is generated by S_4 along with

$$S_1 = \begin{bmatrix} -1 & 2 & 2 & 2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & -1 & 2 & 2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad S_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 2 & 2 & -1 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

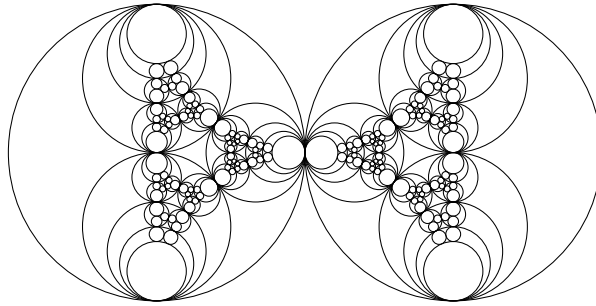


Figure 7. A dual Apollonian packing.

These matrices satisfy $S_i^2 = I$ and $S_i^T Q S_i = Q$ for each i . With this notation, the Apollonian gasket generated by an initial Descartes configuration whose circles have curvature-center matrix M consists of the circles in the configurations of the orbit of M under the left action of \mathcal{A} .

Given a column vector of initial curvatures $(\kappa_1, \kappa_2, \kappa_3, \kappa_4)^T$ that satisfy the Descartes condition, in light of (3) and the above, multiplication by S_i can be viewed as removing curvature κ_i and substituting the curvature of its replacement. Thus the curvatures that occur in an Apollonian gasket with initial curvature vector v are those that occur in the vectors of the orbit of v under the action of the Apollonian group.

One can verify that the matrix $T_i := S_i^T$ corresponds to inversion over the i -th circle of a configuration, that the matrix $D := -\frac{1}{2}Q$ gives $DM(C_1, \dots, C_4) = M(C'_1, \dots, C'_4)$, and that $D = D^{-1} = D^T$. These matrices are related by $S_i D = D T_i$ for each i . As a result, the dual Apollonian group \mathcal{A}^\perp generated by T_1, \dots, T_4 is conjugate to the Apollonian group. The orbit of a configuration under \mathcal{A}^\perp is called a dual Apollonian packing; see Figure 7.

4. An abstract supergasket

Having now reviewed the geometric and algebraic constructions of Apollonian circle packings, we proceed to transpose the algebraic viewpoint; instead of looking at configurations, we will focus on identifying individual circles. From now on, for convenience, we will view (a, b, c, d) both as a point and as a vector. We will also use it to identify a circle: given a configuration with curvature-center matrix M , let (a, b, c, d) be the circle whose curvature-center coordinates are given by the vector $(a, b, c, d)M$.

There are two motivations for this notation. One is to extend Northshield's coordinates for Ford circles. The other is to view the process of generating an Apollonian gasket in an abstract fashion: if M is the curvature-center matrix

of the configuration that generates an Apollonian gasket, then by definition any configuration in the gasket has curvature-center matrix of the form AM , where A is an element of the Apollonian group \mathcal{A} . In particular, $M = IM$, and we can view the rows of the identity matrix I as giving the four original circles, which correspond to the labels $e_1 = (1, 0, 0, 0)$, $e_2 = (0, 1, 0, 0)$, $e_3 = (0, 0, 1, 0)$, and $e_4 = (0, 0, 0, 1)$.

Information contained in these labels can be applied to any gasket by using the corresponding curvature-center matrix. For example, using the curvature-center coordinates of the first four circles in the Ford gasket, the reader can verify that each label (a, b, c, d) with $a + b \neq 0$ corresponds to the circle with

$$x = \frac{b}{a+b}, \quad y = \frac{a+b-c+d}{2(a+b)}, \quad k = 2(a+b), \tag{4}$$

where (x, y) is the center and k is the curvature, while labels of the form $(a, -a, c, d)$ correspond to lines.

While any label (a, b, c, d) corresponds to a circle, which ones give circles in the gasket? This question is equivalent to asking what rows can occur in matrices in \mathcal{A} . If l is a circle in the gasket, then l is a row of some matrix $A \in \mathcal{A}$ and, for any i , we have $AS_i \in \mathcal{A}$. Then lS_i is a row of AS_i , and so lS_i is the label of a circle in the gasket. Since any $A \in \mathcal{A}$ can be written as a word in the S_i , any vector corresponding to the label of a circle in the gasket can be written as $e_i A$ for some $A \in \mathcal{A}$ and some $1 \leq i \leq 4$. Thus the question becomes what are the orbits of the e_i under \mathcal{A} ?

Let

$$f_Q(a, b, c, d) = 2(a^2 + b^2 + c^2 + d^2) - (a + b + c + d)^2.$$

Then $f_Q(e_i) = \langle e_i, e_i \rangle_Q = 1$ for each i . Moreover, since $f_Q(e_i) = 1$ and $\langle uS_i, uS_i \rangle_Q = \langle u, u \rangle_Q$ for each i and every vector u , each label (a, b, c, d) of a circle in the gasket satisfies $f_Q(a, b, c, d) = 1$. Unfortunately, this condition does not characterize the gasket circles.³ One way to discover this is to start plotting integer solutions to $f_Q(a, b, c, d) = 1$ using (4); in doing so, an interesting picture emerges (see Figure 8).

The group \mathcal{A}^S generated by the S_i and the T_i is the super Apollonian group, and an orbit of a configuration under the super Apollonian group is a superpacking or *supergasket* [Graham et al. 2006]. In fact, as we will prove, integer solutions to $f_Q(a, b, c, d) = 1$ correspond bijectively to the circles of any Apollonian supergasket. The rest of this section is devoted to proving this characterization.

Let \mathcal{I} be the set of integer solutions to $f_Q(a, b, c, d) = 1$. Note first that $\langle uS_i, uS_i \rangle_Q = \langle uT_i, uT_i \rangle_Q = \langle u, u \rangle_Q$ for each i and every vector u , so that $\langle e_i A, e_i A \rangle_Q = 1$ for each i and any $A \in \mathcal{A}^S$. Thus each orbit of an e_i is a subset

³Such a condition would be of much interest, and we mention this again as an open problem later.

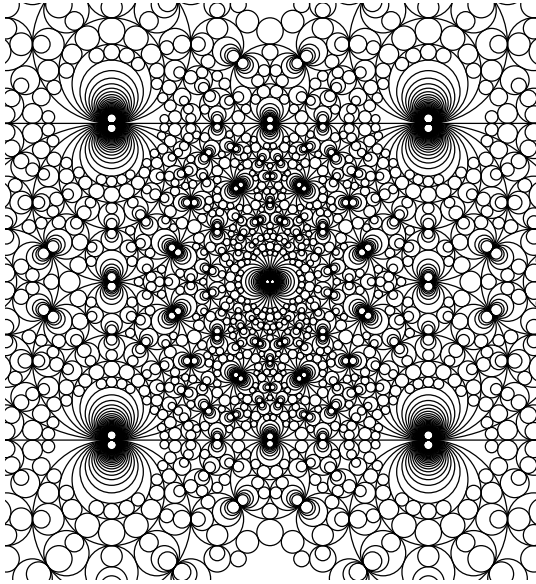


Figure 8. Plot of integer solutions to $f_Q(a, b, c, d) = 1$.

of \mathcal{I} . Our next few results explore properties of \mathcal{I} . One fact we will use repeatedly is that $(a, b, c, d) \in \mathcal{I}$ means

$$a = b + c + d \pm \sqrt{4(bc + bd + cd) + 1}. \tag{5}$$

Lemma 2. *There is no element of \mathcal{I} with two negative coordinates and two positive coordinates.*

Proof. Assume without loss of generality that a and b are negative and that c and d are positive, and rewrite $f_Q(a, b, c, d) = 1$ as

$$(a - b)^2 + (c - d)^2 = 2(a + b)(c + d) + 1. \tag{6}$$

Then the left side is positive but the right is negative, a contradiction. □

If $(a, b, c, d) \in \mathcal{I}$, then $(-a, -b, -c, -d) \in \mathcal{I}$, and they are the same circle but with opposite orientations. Since orientation changes are already present in the curvature-center matrices, they should not be needed in the labels. Let \mathcal{I}^+ be the subset of \mathcal{I} consisting of labels with at least one positive coordinate and at least as many positive coordinates as negative.

Our eventual proof that \mathcal{I}^+ will behave as the abstract supergasket will depend on an algorithm to take any element of \mathcal{I}^+ and produce a series of transformations that will take us back to some e_i . The next four results show that the S and T transformations map \mathcal{I}^+ to itself.

Lemma 3. *Let $(a, b, c, d) \in \mathcal{I}^+$ have no negative entries and let $a = \max\{a, b, c, d\}$. Then $b + c + d < a$. Further, $a < 3(b + c + d)$ unless $(a, b, c, d) = e_1$.*

Proof. If $a \leq b + c + d$, then (5) implies $-(bc + bd + cd) > \frac{1}{2}$, a contradiction. If $a \geq 3(b + c + d)$, then (5) yields $b^2 + c^2 + d^2 \leq \frac{1}{2}$, implying $b = c = d = 0$. \square

Lemma 4. *For $(a, b, c, d) \in \mathcal{I}^+$ with no negative entries and $a = \max\{a, b, c, d\}$, unless $(a, b, c, d) = e_1$, we have*

$$(a', b', c', d') := (a, b, c, d)T_1 \in \mathcal{I}^+ \quad \text{and} \quad a + b + c + d > a' + b' + c' + d' > 0.$$

Proof. Since T_1 only changes a , we know $(a, b, c, d)T_1$ has at most one negative entry. Thus, if $(a, b, c, d) \neq e_1$, then $(a, b, c, d)T_1 \in \mathcal{I}^+$. Further, $a' + b' + c' + d' = 3b + 3c + 3d - a$, so assuming $(a, b, c, d) \neq e_1$, we have $3b + 3c + 3d - a > a - a = 0$. Using $b + c + d < a$,

$$a' + b' + c' + d' - a - b - c - d = 2b + 2c + 2d - 2a > 0. \quad \square$$

Lemma 5. *Let $(a, b, c, d) \in \mathcal{I}^+$ have exactly one negative entry a . Then $a \geq -\frac{1}{6}(b + c + d)$. If $a = -\frac{1}{6}(b + c + d)$ then $(a, b, c, d) = (-1, 2, 2, 2)$.*

Proof. Assume $a \leq -\frac{1}{6}(b + c + d)$. Then (5) implies

$$36 \geq 49(b^2 + c^2 + d^2) - 46(bc + bd + cd).$$

Assume without loss of generality that $d \geq c \geq b \geq 0$. Using that

$$b^2 + c^2 + d^2 - bc - bd - cd = (b - c)^2 + (d - b)(d - c) \geq 0,$$

we have $12 \geq b^2 + c^2 + d^2$. The only such nonnegative values of b, c , and d that admit an a with $f_Q(a, b, c, d) = 1$ are $b = c = d = 2$. \square

Lemma 6. *For $(a, b, c, d) \in \mathcal{I}^+$ with exactly one negative entry a ,*

$$(a', b', c', d') := (a, b, c, d)S_1 \in \mathcal{I}^+ \quad \text{and} \quad a + b + c + d > a' + b' + c' + d' > 0.$$

Proof. Since a is negative, $a' = -a$ is positive. If $(a', b', c', d') \notin \mathcal{I}^+$, then by Lemma 2, each of b', c' , and d' are negative. Thus $b + 2a = b' < 0$, and similarly $c + 2a < 0$ and $d + 2a < 0$. Taken together, $b + c + d + 6a < 0$, a contradiction.

Since $a < 0$, it follows that $a' + b' + c' + d' - a - b - c - d = 4a > 0$. Finally,

$$a' + b' + c' + d' = 5a + b + c + d > 6a + b + c + d > 0. \quad \square$$

Now for the main result that establishes the connection between \mathcal{I}^+ and the action of \mathcal{A}^S .

Lemma 7. *Suppose $l \in \mathcal{I}^+$. There exists an element $A \in \mathcal{A}^S$ and an i such that $l = e_i A$.*

Proof. Since $l = (a_1, a_2, a_3, a_4) \in \mathcal{I}^+$, either it has no negative entries or exactly one negative entry. Consider the operation

$$l \mapsto \begin{cases} IT_i & \text{if } l \text{ has no negative entries and } a_i = \max\{a_1, a_2, a_3, a_4\}, \\ IS_j & \text{if } l \text{ has exactly one negative entry } a_j. \end{cases}$$

By Lemmas 4 and 6, repeated application of this operation will eventually result in e_i for some i and we will have $lA = e_i$ for some $A \in \mathcal{A}^S$. Since each T_i and S_i are invertible, $l = e_i A^{-1}$. □

Conversely, for any $A \in \mathcal{A}^S$ and any i , we have $e_i A \in \mathcal{A}^S$, establishing our bijection.

Theorem 8. *The circles of an Apollonian supergasket are in one-to-one correspondence with \mathcal{I}^+ .*

If $l = e_i A$ as in Lemma 7, then the three circles $e_j A$ for $j \neq i$ form a configuration with l , and we can call them the “parents” of l . From (6), the elements of \mathcal{I} must have exactly one odd entry, and one can verify the location of this entry is not altered by replacement or inversion. Thus the odd entry provides a quick indicator of which e_i will be obtained by the procedure of Lemma 7.

Since duality D preserves the Q -inner product, the labels (a, b, c, d) of dual circles also satisfy $f_Q(a, b, c, d) = 1$, but the one odd entry of elements of \mathcal{I} means that the elements of $2\mathcal{I}D$ are all odd integers. Results similar to Lemmas 2, 3, and 5 hold for dual circles, and thus a procedure similar to that of Lemma 7 can return a dual circle to one of the original four dual circles: $(-\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ or a permutation thereof.

5. Label operations

Having now defined our abstract supergasket as the set \mathcal{I}^+ , we can begin to put it to use. We are particularly interested in properties shared by all gaskets. As we will see in this section, the labels give a simple way to identify individual circles, but they can also be combined to give simple computations for the configuration operations. As a first example, the next theorem follows directly from analyzing the entries of the S_i .

Theorem 9. *Let C_1, \dots, C_4 be the circles of a Descartes configuration with labels c_1, \dots, c_4 . Let C_5 be the replacement of C_4 , let C'_j be the inversion of C_j , $2 \leq j \leq 4$, over C_1 , and let c_5 and c'_j denote the corresponding labels. Using entrywise operations, $c_5 = 2c_1 + 2c_2 + 2c_3 - c_4$ and $c'_j = 2c_1 + c_j$.*

A key fact is that, using duality and as witnessed by $S_i D = D T_i$ for each i , replacement can be viewed as inversion and inversion can be viewed as replacement. As an example of an application, for any circle C in the plane and any Descartes configuration with curvature-center matrix M , let I_C be the operation of inversion

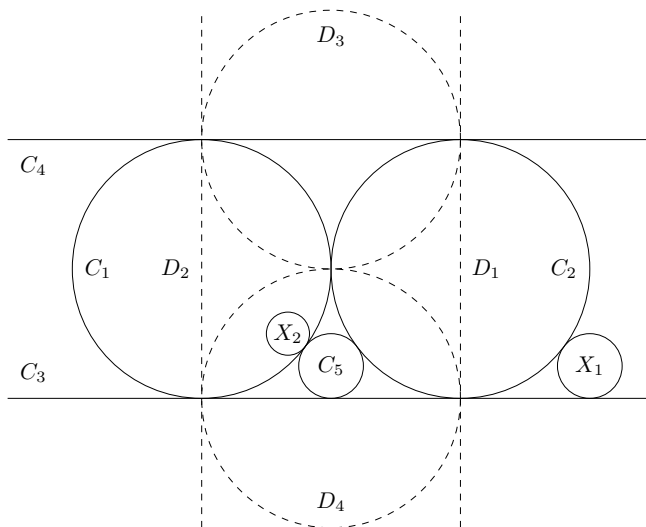


Figure 9. Circle X_1 is the inversion of C_5 over D_1 , and X_2 is the inversion of C_5 over C_1 .

over C . If, for some i , the intersection of the interior of C with the interior of any circle represented by M or $S_i M$ is empty, then $S_i I_C M = I_C S_i M$, and the similar results hold for T_i and for duality D . To see this, recall that the replacement of a circle determines a unique circle tangent to the other three in the original configuration. Inversion preserves tangency, and the unique circle tangent to three of $I_C M$ must be the inversion of the unique circle tangent to the corresponding three of M . Duality is similarly uniquely defined by the points of intersection which preserve their status under inversion. This view can help us to understand the action of an individual S_i or T_i on a given label, since multiplication of a label vector on the right corresponds to “premultiplication” on the left of the matrix M for a configuration.

Theorem 10. *Multiplication of a label vector on the right by T_i corresponds to inversion over the i -th circle of the original configuration, while multiplication on the right by S_i corresponds to inversion over the i -th dual circle.*

For example, using (4), the label $(1, 0, 0, 0)$ corresponds to the circle with center $(0, \frac{1}{2})$ and curvature 2, called C_1 in Figure 9, and $(0, 1, 0, 0)$, $(0, 0, 1, 0)$, and $(0, 0, 0, 1)$ correspond to C_2 , C_3 , and C_4 , respectively. The dual circles are the D_i . According to Theorem 9, C_5 has label $(2, 2, 2, -1)$. According to Theorem 10, for example, $(2, 2, 2, -1)S_1 = (-2, 6, 6, 3)$ gives circle X_1 , which is the inversion of C_5 over D_1 , and $(2, 2, 2, -1)T_1 = (4, 2, 2, -1)$ gives circle X_2 , which is the inversion of C_5 over C_1 .

6. An inner product

Curvature-center coordinate vectors take on another meaning when viewed in \mathbb{R}^4 with the indefinite inner product $\langle \cdot, \cdot \rangle_G$ given by the matrix

$$G = \frac{1}{2} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{bmatrix}.$$

For circles C_1 and C_2 that are not lines, let d be the distance between their centers and let r_1 and r_2 be their respective radii. If C_1 and C_2 intersect at an angle θ , then $d^2 = r_1^2 + r_2^2 - 2r_1r_2 \cos \theta$. Define a quantity A for C_1 and C_2 as $2Ar_1r_2 = d^2 - r_1^2 - r_2^2$ [Kotlov et al. 1997]. A then generalizes the intersection angle to any pair of circles. Moreover, if v_1 and v_2 are the curvature-center coordinate vectors of C_1 and C_2 , respectively, then $A = \langle v_1, v_2 \rangle_G = v_1 G v_2^T$. For two circles C_1 and C_2 (including lines), letting $\langle C_1, C_2 \rangle_G$ be the G -inner product of their curvature-center vectors, we get the following characterization:

$\langle C_1, C_2 \rangle_G$	C_1 and C_2
-1	are internally tangent
1	are externally tangent
0	are mutually orthogonal
$-\cos \alpha$	intersect at angle α
< -1	are disjoint, one inside the other
> 1	are disjoint, outside each other

In general, given four circles C_1, \dots, C_4 with curvature-center coordinate vectors v_1, \dots, v_4 , Jerzy Kocik [2007] defines their *configuration matrix* $F = F(C_1, \dots, C_4)$ to be the Gram matrix of the vectors v_1, \dots, v_4 with respect to $\langle \cdot, \cdot \rangle_G$; that is, $F_{ij} = \langle v_i, v_j \rangle_G$. Thus if M is the curvature-center matrix for C_1, \dots, C_4 , then $F = MGM^T$.

For a (Descartes) configuration, the configuration matrix F is $-Q$. In that case, F is invertible, thus so is M , and $F = MGM^T$ if and only if $M^T F^{-1} M = G^{-1}$. The inverses of F and G are also related to previously defined matrices: $G^{-1} = -\frac{1}{4}W$ and $F^{-1} = -\frac{1}{4}Q$.

From the above, if M is the curvature-center matrix of a Descartes configuration, $MGM^T = -Q$. Thus for labels u and v , we have $\langle u, v \rangle_Q = -\langle uM, vM \rangle_G$, so that Q -inner products of our label vectors also give the geometric relationships between the circles they represent. For example, letting $\langle C_1, C_2 \rangle_Q$ be the Q -inner product of the labels of circles C_1 and C_2 , we have the following theorem.

Theorem 11. *Circles C_1 and C_2 are externally tangent, mutually orthogonal, or internally tangent if and only if $\langle C_1, C_2 \rangle_Q$ is $-1, 0, or 1, respectively.$*

Viewing the circles as vectors suggests additional constructions, including one that resembles a Householder transformation:⁴ Let C be any circle in a superpacking and let c be its label. For other labels d , consider the map $d \mapsto d(I - 2Qc^T c)$ (with labels used as vectors). Since C is internally tangent to itself, $\langle C, C \rangle_Q = cQc^T = 1$ and this map is an involution. Moreover, for any circle C' tangent to C , from Theorem 11 we have $\langle C, C' \rangle_Q = -1$, so that $c' \mapsto c' + 2c$. From Theorem 9, this map inverts the circles tangent to C over C . Finally, every other circle in the supergasket can be obtained via replacement and/or duality and we saw earlier that those operations commute with inversion over C .

Theorem 12. *If c and d are circles in the abstract superpacking, then $d(I - 2Qc^T c)$ is the inversion of d over c .*

Note that by computing $(I - 2Qe_i^T e_i)$ for $i \in \{1, \dots, 4\}$, Theorem 12 also provides another justification for part of Theorem 10.

7. Curvatures

We return now to the fascinating problem mentioned at the start: given four original integer curvatures, which other curvatures can or will occur? Certain conditions modulo 24 are known [Graham et al. 2003], and recent progress has been made in the form of a positive density theorem [Bourgain and Fuchs 2011] and a local-global theorem [Bourgain and Kontorovich 2014]. Our labels can provide an analysis similar to the proof of the positive density theorem, which involves looking at the curvatures of circles tangent to a given circle.

In the proof of the positive density theorem, if a, b, c , and d are the curvatures of the first four circles, then the set of curvatures of the circles tangent to the circle C_1 of curvature a involves the quadratic form $f(x, y) = Ax^2 + 2Bxy + Cy^2$, where $A = a + b$, $B = \frac{1}{2}(a + b + d - c)$, and $C = a + d$. In particular, the set of curvatures of the circles tangent to C_1 is shown to contain the set $\{f(x, y) - a : \gcd(x, y) = 1\}$. For our approach, notice that the Ford circles are the circles tangent to one of the four original circles in the Ford gasket (the x -axis). Our labels extend Northshield's [2015] in that the abstract Ford circles are (s, t, u, v) , where $\gcd(s, t, u) = 1$ and $s^2 + t^2 + u^2 = (s + t + u)^2$. In particular, using Northshield's ideas, the abstract Ford circle labels can be parametrized as

$$(x(x + y), y(x + y), x^2 + xy + y^2 - 1, -xy)$$

with $\gcd(x, y) = 1$. Thus, if a, b, c , and d are the initial curvatures of a gasket, then

$$(x^2 + xy + y^2 - 1, x(x + y), -xy, y(x + y))$$

⁴A Householder transformation of a vector is the result of multiplication by a matrix of the form $I - vv^T$, where I is an identity matrix and v is a column vector of the appropriate size.

has curvature

$$a(x^2 + xy + y^2 - 1) + b(x(x + y)) + c(-xy) + d(y(x + y)) = f(x, y) - a.$$

Equation (6) also gives some information about the set of curvatures of the Ford supergasket since $2(a + b)$ is the curvature of the circle (a, b, c, d) . In particular, given a desired curvature κ , the equations

$$2(a + b) = \kappa, \quad a - b = y_1, \quad c - d = y_2, \quad \text{and} \quad c + d = y_3$$

provide a connection to the solutions of the equation $y_1^2 + y_2^2 = \kappa y_3 + 1$. Recalling Fermat's result that any number of the form pq^2 , where the prime factorization of p consists of primes that are congruent to 1 modulo 4, can be written as the sum of two perfect squares gives a quick way to see that every integer occurs as a curvature in the Ford supergasket.

Ideally, we could characterize the subset of supergasket labels that form a gasket and find a parametrization using that characterization. Suppose $f_Q(a, b, c, d) = 1$ and d is odd. Then $4(ab + ac + bc) + 1$ is a perfect square, say m^2 , so $4(ab + ac + bc) = m^2 - 1$ and m must be odd. Thus $ab + ac + bc = n(n - 1)$ for some integer n . Conversely, if $ab + ac + bc = n(n - 1)$, then $4(ab + ac + bc) + 1$ is a perfect square. Perhaps there exists a simple characterization of the n that occur in this fashion.

Acknowledgments

This work was supported by NSF grant DMS 11-56890 and Central Michigan University. We would also like to thank Yeon Kim and Sivaram Narayan for helpful discussions and the referees for their comments.

References

- [Bourgain 2012] J. Bourgain, "Integral Apollonian circle packings and prime curvatures", *J. Anal. Math.* **118**:1 (2012), 221–249. MR Zbl
- [Bourgain and Fuchs 2011] J. Bourgain and E. Fuchs, "A proof of the positive density conjecture for integer Apollonian circle packings", *J. Amer. Math. Soc.* **24**:4 (2011), 945–967. MR Zbl
- [Bourgain and Kontorovich 2014] J. Bourgain and A. Kontorovich, "On the local-global conjecture for integral Apollonian gaskets", *Invent. Math.* **196**:3 (2014), 589–650. MR Zbl
- [Fuchs 2013] E. Fuchs, "Counting problems in Apollonian packings", *Bull. Amer. Math. Soc. (N.S.)* **50**:2 (2013), 229–266. MR Zbl
- [Fuchs and Sanden 2011] E. Fuchs and K. Sanden, "Some experiments with integral Apollonian circle packings", *Exp. Math.* **20**:4 (2011), 380–399. MR Zbl
- [Graham et al. 2003] R. L. Graham, J. C. Lagarias, C. L. Mallows, A. R. Wilks, and C. H. Yan, "Apollonian circle packings: number theory", *J. Number Theory* **100**:1 (2003), 1–45. MR Zbl
- [Graham et al. 2005] R. L. Graham, J. C. Lagarias, C. L. Mallows, A. R. Wilks, and C. H. Yan, "Apollonian circle packings: geometry and group theory, I: The Apollonian group", *Discrete Comput. Geom.* **34**:4 (2005), 547–585. MR Zbl

- [Graham et al. 2006] R. L. Graham, J. C. Lagarias, C. L. Mallows, A. R. Wilks, and C. H. Yan, “Apollonian circle packings: geometry and group theory, II: Super-Apollonian group and integral packings”, *Discrete Comput. Geom.* **35**:1 (2006), 1–36. MR Zbl
- [Guettler and Mallows 2010] G. Guettler and C. Mallows, “A generalization of Apollonian packing of circles”, *J. Comb.* **1**:1 (2010), 1–27. MR Zbl
- [Hirst 1967] K. E. Hirst, “The Apollonian packing of circles”, *J. London Math. Soc.* **42** (1967), 281–291. MR Zbl
- [Kocik 2007] J. Kocik, “A theorem on circle configurations”, preprint, 2007. arXiv
- [Kotlov et al. 1997] A. Kotlov, L. Lovász, and S. Vempala, “The Colin de Verdière number and sphere representations of a graph”, *Combinatorica* **17**:4 (1997), 483–521. MR Zbl
- [Lagarias et al. 2002] J. C. Lagarias, C. L. Mallows, and A. R. Wilks, “Beyond the Descartes circle theorem”, *Amer. Math. Monthly* **109**:4 (2002), 338–361. MR Zbl
- [Northshield 2015] S. Northshield, “Ford circles and spheres”, preprint, 2015. arXiv
- [Oh 2014] H. Oh, “Apollonian circle packings: dynamics and number theory”, *Jpn. J. Math.* **9**:1 (2014), 69–97. MR Zbl
- [Sarnak 2011] P. Sarnak, “Integral Apollonian packings”, *Amer. Math. Monthly* **118**:4 (2011), 291–306. MR Zbl

Received: 2016-02-16 Revised: 2016-03-16 Accepted: 2016-03-19

isabel.corona@colorado.edu	<i>Department of Mathematics, University of Colorado Boulder, Boulder, CO 80309, United States</i>
chjohnson@middlebury.edu	<i>Department of Mathematics, Middlebury College, 14 Old Chapel Road, Middlebury, VT 05753, United States</i>
lhm@ams.org	<i>Mathematical Reviews, American Mathematical Society, 416 4th Street, Ann Arbor, MI 48103, United States</i>
doconnel@haverford.edu	<i>Mathematics Department, Haverford College, 370 Lancaster Avenue, Haverford, PA 19041, United States</i>

Guidelines for Authors

Submissions in all mathematical areas are encouraged. All manuscripts accepted for publication in *Involve* are considered publishable in quality journals in their respective fields, and include a minimum of one-third student authorship. Submissions should include substantial faculty input; faculty co-authorship is strongly encouraged. Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use L^AT_EX but submissions in other varieties of T_EX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibT_EX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

involve

2017 vol. 10 no. 2

Stability analysis for numerical methods applied to an inner ear model	181
KIMBERLEY LINDENBERG, KEES VUIK AND PIETER W. J. VAN HENGEL	
Three approaches to a bracket polynomial for singular links	197
CARMEN CAPRAU, ALEX CHICHESTER AND PATRICK CHU	
Symplectic embeddings of four-dimensional ellipsoids into polydiscs	219
MADELEINE BURKHART, PRIERA PANESCU AND MAX TIMMONS	
Characterizations of the round two-dimensional sphere in terms of closed geodesics	243
LEE KENNARD AND JORDAN RAINONE	
A necessary and sufficient condition for coincidence with the weak topology	257
JOSEPH CLANIN AND KRISTOPHER LEE	
Peak sets of classical Coxeter groups	263
ALEXANDER DIAZ-LOPEZ, PAMELA E. HARRIS, ERIK INSKO AND DARLEEN PEREZ-LAVIN	
Fox coloring and the minimum number of colors	291
MOHAMED ELHAMDADI AND JEREMY KERR	
Combinatorial curve neighborhoods for the affine flag manifold of type A_1^1	317
LEONARDO C. MIHALCEA AND TREVOR NORTON	
Total variation based denoising methods for speckle noise images	327
ARUNDHATI BAGCHI MISRA, ETHAN LOCKHART AND HYEONA LIM	
A new look at Apollonian circle packings	345
ISABEL CORONA, CAROLYNN JOHNSON, LON MITCHELL AND DYLAN O'CONNELL	



1944-4176(2017)10:2;1-B