

# involve

a journal of mathematics

## Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

Colin Adams	Suzanne Lenhart
John V. Baxley	Chi-Kwong Li
Arthur T. Benjamin	Robert B. Lund
Martin Bohner	Gaven J. Martin
Nigel Boston	Mary Meyer
Amarjit S. Budhiraja	Emil Minchev
Pietro Cerone	Frank Morgan
Scott Chapman	Mohammad Sal Moslehian
Jem N. Corcoran	Zuhair Nashed
Toka Diagana	Ken Ono
Michael Dorff	Timothy E. O'Brien
Sever S. Dragomir	Joseph O'Rourke
Behrouz Emamizadeh	Yuval Peres
Joel Foisy	Y.-F. S. Pétermann
Errin W. Fulp	Robert J. Plemmons
Joseph Gallian	Carl B. Pomerance
Stephan R. Garcia	Bjorn Poonen
Anant Godbole	James Propp
Ron Gould	Józeph 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
David Larson	



# involve

msp.org/involve

## EDITORS

### MANAGING EDITOR

Kenneth S. Berenhaut, Wake Forest University, USA, berenhks@wfu.edu

### BOARD OF EDITORS

Colin Adams	Williams College, USA colin.c.adams@williams.edu	David Larson	Texas A&M University, USA larson@math.tamu.edu
John V. Baxley	Wake Forest University, NC, USA baxley@wfu.edu	Suzanne Lenhart	University of Tennessee, USA lenhart@math.utk.edu
Arthur T. Benjamin	Harvey Mudd College, USA benjamin@hmc.edu	Chi-Kwong Li	College of William and Mary, USA ckli@math.wm.edu
Martin Bohner	Missouri U of Science and Technology, USA bohner@mst.edu	Robert B. Lund	Clemson University, USA lund@clemson.edu
Nigel Boston	University of Wisconsin, USA boston@math.wisc.edu	Gaven J. Martin	Massey University, New Zealand g.j.martin@massey.ac.nz
Amarjit S. Budhiraja	U of North Carolina, Chapel Hill, USA budhiraj@email.unc.edu	Mary Meyer	Colorado State University, USA meyer@stat.colostate.edu
Pietro Cerone	La Trobe University, Australia P.Cerone@latrobe.edu.au	Emil Minchev	Ruse, Bulgaria eminchev@hotmail.com
Scott Chapman	Sam Houston State University, USA scott.chapman@shsu.edu	Frank Morgan	Williams College, USA frank.morgan@williams.edu
Joshua N. Cooper	University of South Carolina, USA cooper@math.sc.edu	Mohammad Sal Moselehian	Ferdowsi University of Mashhad, Iran moslehian@ferdowsi.um.ac.ir
Jem N. Corcoran	University of Colorado, USA corcoran@colorado.edu	Zuhair Nashed	University of Central Florida, USA znashed@mail.ucf.edu
Toka Diagana	Howard University, USA tdiagana@howard.edu	Ken Ono	Emory University, USA ono@mathcs.emory.edu
Michael Dorff	Brigham Young University, USA mdorff@math.byu.edu	Timothy E. O'Brien	Loyola University Chicago, USA tobrie1@luc.edu
Sever S. Dragomir	Victoria University, Australia sever@matilda.vu.edu.au	Joseph O'Rourke	Smith College, USA orourke@cs.smith.edu
Behrouz Emamizadeh	The Petroleum Institute, UAE bemamizadeh@pi.ac.ae	Yuval Peres	Microsoft Research, USA peres@microsoft.com
Joel Foisy	SUNY Potsdam foisyjs@potsteam.edu	Y.-F. S. Pétermann	Université de Genève, Switzerland petermann@math.unige.ch
Errin W. Fulp	Wake Forest University, USA fulp@wfu.edu	Robert J. Plemmons	Wake Forest University, USA rplemmons@wfu.edu
Joseph Gallian	University of Minnesota Duluth, USA jgallian@d.umn.edu	Carl B. Pomerance	Dartmouth College, USA carl.pomerance@dartmouth.edu
Stephan R. Garcia	Pomona College, USA stephan.garcia@pomona.edu	Vadim Ponomarenko	San Diego State University, USA vadim@sciences.sdsu.edu
Anant Godbole	East Tennessee State University, USA godbole@etsu.edu	Bjorn Poonen	UC Berkeley, USA poonen@math.berkeley.edu
Ron Gould	Emory University, USA rg@mathcs.emory.edu	James Propp	U Mass Lowell, USA jpropp@cs.uml.edu
Andrew Granville	Université Montréal, Canada andrew@dms.umontreal.ca	József H. Przytycki	George Washington University, USA przytyck@gwu.edu
Jerrold Griggs	University of South Carolina, USA griggs@math.sc.edu	Richard Rebarber	University of Nebraska, USA rrebarbe@math.unl.edu
Sat Gupta	U of North Carolina, Greensboro, USA sngupta@uncg.edu	Robert W. Robinson	University of Georgia, USA rwr@cs.uga.edu
Jim Haglund	University of Pennsylvania, USA jhaglund@math.upenn.edu	Filip Saidak	U of North Carolina, Greensboro, USA f_saidak@uncg.edu
Johnny Henderson	Baylor University, USA johnny_henderson@baylor.edu	James A. Sellers	Penn State University, USA sellersj@math.psu.edu
Jim Hoste	Pitzer College jhoste@pitzer.edu	Andrew J. Sterge	Honorary Editor andy@ajsterge.com
Natalia Hritonenko	Prairie View A&M University, USA nahritonenko@pvamu.edu	Ann Trenk	Wellesley College, USA atrenk@wellesley.edu
Glenn H. Hurlbert	Arizona State University, USA hurlbert@asu.edu	Ravi Vakil	Stanford University, USA vakil@math.stanford.edu
Charles R. Johnson	College of William and Mary, USA crjohnso@math.wm.edu	Antonia Vecchio	Consiglio Nazionale delle Ricerche, Italy antonia.vecchio@cnr.it
K. B. Kulasekera	Clemson University, USA kk@ces.clemson.edu	Ram U. Verma	University of Toledo, USA verma99@msn.com
Gerry Ladas	University of Rhode Island, USA gladas@math.uri.edu	John C. Wierman	Johns Hopkins University, USA wierman@jhu.edu
		Michael E. Zieve	University of Michigan, USA zieve@umich.edu

## PRODUCTION


Silvio Levy, Scientific Editor

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2015 is US \$140/year for the electronic version, and \$190/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<sup>®</sup> from Mathematical Sciences Publishers.

PUBLISHED BY

 **mathematical sciences publishers**  
nonprofit scientific publishing

<http://msp.org/>

© 2015 Mathematical Sciences Publishers

# Enhancing multiple testing: two applications of the probability of correct selection statistic

Erin Irwin and Jason Wilson

(Communicated by Timothy O'Brien)

The calculation of the probability of correct selection (PCS) shows how likely it is that the populations chosen as “best” truly are the top populations, according to a well-defined standard. PCS is useful for the researcher with limited resources or the statistician attempting to test the quality of two different statistics. This paper explores the theory behind two selection goals for PCS,  $G$ -best and  $d$ -best, and how they improve previous definitions of PCS for massive datasets. This paper also calculates PCS for two applications that have already been analyzed by multiple testing procedures in the literature. The two applications are in neuroimaging and econometrics. It is shown through these applications that PCS not only supports the multiple testing conclusions but also provides further information about the statistics used.

## 1. Introduction

Because of the advancements in technology and science, a new development in statistics must involve correctly and usefully analyzing massive datasets. With internet applications and financial data, there can be as many as ten million populations to analyze, and sometimes more. Statisticians have developed methods such as family-wide error control and the false discovery rate to deal with the multiple testing problem — the problem of finding too many false positives when testing  $k$  hypotheses simultaneously. This paper deals instead with ranking and selection methodology, which is a separate branch of statistics that has also been expanded to apply to massive datasets.

Ranking and selection methodology (RSM) is a well-defined system of ranking a set of populations based on sample data and selecting those that are “best”. In laboratory research, resources are always limited. A scientist may want to know which of 10,000 genes available will provide the most information, to avoid studying

---

*MSC2010:* 46N30, 47N30.

*Keywords:* probability of correct selection (PCS),  $d$ -best,  $G$ -best, ranking and selection, neuroimaging, econometrics.

all of them. Similarly, no one can invest in every company on the stock market, and so an investor only wants to know which ones will make the most money. In these two cases, *best* can be defined as the highest expression levels or the highest average returns on investment, respectively. Traditional hypothesis tests are not meant for ranking and selection purposes. Instead, one can calculate the probability of correct selection (PCS) to evaluate a chosen set of populations and see if the best have actually been chosen.

As with multiple testing procedures, PCS has evolved in the last century from being accurate with large datasets ( $\approx 10$  samples) to being accurate with massive datasets. The two previous methods of ranking and/or selecting the best populations are the indifference zone method (IZ), originated by Robert Bechhofer [1954], and the subset selection method (SS), originated by Shanti S. Gupta [1956]. More recently there have been improvements in PCS for massive datasets by Cui and Wilson [2008] in the form of  $G$ -best and  $d$ -best selection. In this paper, we explore both the theory and some applications of this improved method of calculating PCS.

Specifically, we look at the definitions of  $G$ -best and  $d$ -best selection, the use of index sets in those definitions, and the use of each selection goal. We also apply PCS to a neuroimaging dataset and an econometrics dataset. We find that PCS supports the results found using multiple testing procedures with the neuroimaging application. In addition, PCS provides us with a measure of how accurate our choice of the best populations was. We were able to find the probability that the populations we chose as best based on sample data actually were the best populations. The same information was found for the econometrics data, which is measured by two statistics. PCS was easily adapted for both statistics. These applications show the usefulness of PCS and how it can be applied generally.

The remainder of this paper is organized as follows. Section 2 gives account of the theory behind PCS and  $G$ -best and  $d$ -best selection. Two applications of PCS to datasets already analyzed by multiple testing procedures are given in Section 3. Finally, we draw conclusions in Section 4.

## 2. $G$ -best and $d$ -best selection

**2.1. Introduction.** This section describes the mathematical theory behind  $G$ -best and  $d$ -best selection. The purpose of ranking and selection methodology is ultimately to choose the top  $t$  populations for some specified  $t$ . To do this, we first look at how to denote the ranking of both population parameters and sample statistics. We also use sets of indices to make the definitions of  $G$ -best and  $d$ -best selection more compact. The notation is somewhat subtle, but necessary, and is covered in Section 2.2. Furthermore, in Section 2.3, we define both  $G$ -best and  $d$ -best selection in terms of index sets, and describe how they each meet different needs of

Population $i$	1	2	3	4	5
Sample mean $Y_i$	2.97	1.26	2.90	3.58	1.36
True mean $\theta_i$	2.30	1.70	2.50	4.20	2.50

**Table 1.** Each column of this table of example statistics shows information about a hypothetical population. The populations are numbered  $i = 1, \dots, 5$ . All of this information will be used to illustrate the notation for PCS.

a researcher. We formally state how to calculate PCS in Section 2.4. Finally, we note the improvements these selection goals make for analyzing massive datasets in Section 2.6.

**2.2. Notation.** For clarification, we will use the following example throughout this section. Suppose we know the true means from five populations of interest. We also have taken samples from each population, and have calculated each sample mean. Our example data is given in Table 1.

With this in mind, consider  $k$  populations, each with the same cumulative distribution function (CDF), except with varying location parameter  $\theta_i, i = 1, \dots, k$ . In the example, we have  $k = 5$ . Let  $\theta = (\theta_1, \dots, \theta_k)$  be the vector of these parameters. In Table 1,  $\theta_1$  would then be equal to 2.30. We are really interested in the order of the parameters, and so also have a numbering system for rank. Let  $\theta_{(1)} \leq \dots \leq \theta_{(k)}$  be the ordered parameters of  $\theta$ . For example, in the table,  $\theta_{(1)} = 1.70$ , while  $\theta_{(5)} = 4.20$ .

Sometimes a researcher is interested in the largest statistics, and sometimes the smallest. The definition of the best populations must be defined explicitly for each application. Without loss of generality we assume in this paper that the best population has the largest statistic. What we are ultimately trying to find, then, is the population with the top  $t$  parameters,  $\theta_{(k-t+1)}, \dots, \theta_{(k)}$ , or the top  $t$  parameters themselves. For example, if we want the top  $t = 3$  means from our example, we would want  $\theta_{(5-3+1)}, \dots, \theta_{(5)}$ , or  $\theta_{(3)}, \theta_{(4)}$ , and  $\theta_{(5)}$ .

Because the top parameters are assumed to be unknown, we must pick a statistic  $Y$  to estimate the unknown population parameter. Each statistic will have a continuous CDF.  $Y_i$  denotes the particular statistic of the  $i$ -th population. If we are interested in the usual mean, for example, let the statistic  $Y$  denote the mean, so that

$$\begin{aligned}
 Y_2 &= Y(X_{2,1}, X_{2,2}, X_{2,3}) \\
 &= (.75 + 1.78 + 1.25)/3 = 1.26,
 \end{aligned}$$

where  $X_{2,1}, X_{2,2}, \dots$  denote particular observations from the second population.

To order the sample statistics, we use the notation  $Y_{[i]}$  to indicate that  $Y_{[1]} \leq Y_{[2]} \leq \dots \leq Y_{[k]}$ . On the other hand, we denote by  $Y_{(i)}$  the sample statistic that

Population $i$	2	5	3	1	4
Sample mean	$Y_{[1]} = 1.26$	$Y_{[2]} = 1.36$	$Y_{[3]} = 2.90$	$Y_{[4]} = 2.97$	$Y_{[5]} = 3.58$
True mean	$\theta_{(1)} = 1.70$	$\theta_{(3)} = 2.50$	$\theta_{(3)} = 2.50$	$\theta_{(2)} = 2.30$	$\theta_{(4)} = 4.20$

**Table 2.** This table contains the same information as Table 1, but with the technical notation for PCS added. It is also now sorted by sample mean from lowest to highest.

is drawn from the same population as the ordered parameter  $\theta_{(i)}$ . With Table 1, then,  $Y_{[4]} = Y_{(2)} = 2.97$ , because 2.97 is the fourth largest statistic, but it was from the population with  $\theta_{(2)}$ . With this notation, we can label our data, illustrated in Table 2.

To choose which populations we should assert to be the top  $t$ , we use the top  $t$  statistics. The way we will notate correct selection is using index notation. Let  $s$  be the set of indices of the top  $t$  statistics. For example, if  $t = 1$ , then  $Y_{[5]}$  is the top statistic, but it comes from population  $i = 4$ . So  $s$  in this case would be  $s = \{4\}$ . Then let  $A_t$  be the set of indices of the top  $t$  population parameters. In our case,  $\theta_{(4)}$  is the highest, and also comes from population  $i = 4$ . Therefore,  $A_t = \{4\}$  in this case as well. Rule  $R$  resulting in a correct selection is denoted by

$$CS_t = \{s = A_t\}.$$

Our example would yield a correct selection, then, since the sets  $s$  and  $A_t$  are equal.

It is important to note here that if two population parameters are equal ( $\theta_i = \theta_j$  for some  $i \neq j$ ), both are ranked equally, as we have done in Table 2. In past selection methods, if more than one parameter was equal in value, only one would be randomly chosen and asserted as the correct selection. This may significantly reduce the value of PCS in an unnecessary manner. We will handle this situation in this paper similarly to Cui and Wilson [2008]. If population parameters  $\theta_i$  and  $\theta_j$  are equal, then  $A_t = \{i\}$  or  $\{j\}$ . In other words, if  $t = 1$  and  $\theta_i = \theta_j$  are the top ranked populations, then either  $s = \{i\}$  or  $s = \{j\}$  will result in a correct selection. In our example, if  $t = 2$ , then  $A_t = \{4, 3\}$  or  $A_t = \{4, 5\}$  because 4.20 and 2.50 are the two largest values among the parameters. Therefore  $s = \{4, 3\}$  or  $s = \{4, 5\}$  would both result in a correct selection. This generalizes to handle more than two populations with the same values.

**2.3. *G*-best and *d*-best selection.** All of this notation is the set-up for the two selection goals we discuss in this paper: *G*-best and *d*-best selection, as defined by Cui and Wilson [2008]. These are two different ways to define whether the populations that we choose as best really are the best. One can use these methods before an experiment to determine how many subjects to study in order to ensure the

selection of the top populations. After an experiment, one can use these methods to calculate the probability that the researcher has found the actual top  $t$  populations.

If our previous toy example was an actual experiment, we might have a need to find the top, say, one statistic, but we have the resources to study two. We would then use  $G$ -best selection, where we choose a fixed amount of populations,  $t + G$ , that contains the top  $t$  statistics. On the other hand, if we simply needed the populations to be within a certain threshold of quality, we would use  $d$ -best selection. In  $d$ -best selection, we are finding a random number of populations, say  $r$ , which contains populations that are within a certain distance  $d$  from the top  $t$  populations. The number  $r$  is determined by an interval of prespecified length  $d$ .

**Definition 2.1.** Let  $s$  be the set of the indices corresponding to the top  $t + G$  statistics for some prespecified  $G$ . Let  $A_t$  be a set of indices of the top  $t$  parameters. Then

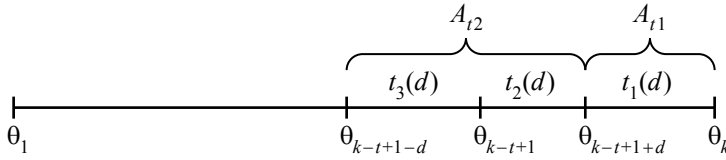
$$CS_{G,t} = \{A_t \subseteq s\}.$$

A set  $s$  that satisfies  $CS_{G,t}$  is called  $G$ -best, and the probability that we have chosen a  $G$ -best set is denoted by  $P(CS_{G,t})$ .

Note that in the case that every population parameter is unique, then  $A_t$  will also be unique. If two or more population parameters are equal, then  $A_t$  may not be unique. Thus, the definition only calls for a possible set of  $A_t$  to be a subset of  $s$  in order to satisfy  $CS_{G,t}$ .

For example, let  $t = 2$  and let  $G = 1$ . We will choose  $t + G = 3$  populations that we assert to contain the top two populations. We would then choose, from Table 2,  $Y_4 = Y_{[5]} = 3.58$ ,  $Y_1 = Y_{[4]} = 2.97$  and  $Y_3 = Y_{[3]} = 2.90$  as our top three statistics, to make  $s = \{1, 3, 4\}$ .  $A_t$  would be  $A_t = \{4, 3\}$  or  $\{4, 5\}$ . In this case, since one possible  $A_t$  is contained in  $s$ , we have chosen correctly.

With this definition, a set is not  $G$ -best unless we have actually chosen the top  $t$  statistics. On the other hand, instead of only choosing  $t$  statistics to work with, we are choosing  $t + G$  for some prespecified  $G$ . Thus, the  $G$  parameter allows one to control the minimum proportion of best populations in the correct selection. For example, selecting the top 20 out of 20 voxel clusters in a neuroimaging scan might be highly unlikely. In this scenario we would have  $t = 20$  and  $G = 0$ . This does not allow for any of the chosen populations to be wrong. However, suppose we can determine that having 90% of the populations actually being the best is allowable. In this case,  $t = 18$  and  $G = 2$ , and the top 18 out of 20 might have a reasonable chance of actually being correct. It may be a low  $P(CS_{G,t})$ , but a reasonable chance is still an improvement. The point of ranking and selection procedures is to narrow down the populations to the best ones, and controlling the proportion of top populations among a group of populations does this.



**Figure 1.** The labels  $t_1(d)$ ,  $t_2(d)$ , and  $t_3(d)$  denote the number of population parameters in their respective intervals. Note that  $t_1(d) + t_2(d) = t$ .  $A_{t_1}$  and  $A_{t_2}$  are the sets of indices of the populations with values in their respective intervals.

**Definition 2.2.** Let  $s$  be the set of the indices corresponding to the top  $t$  statistics. Let  $A_{t_1}$  be the set of indices of the parameters in the interval  $(\theta_{(k-t+1)} + d, \theta_{(k)})$ , and  $A_{t_2}$  the set of indices of the parameters in the interval  $[\theta_{(k-t+1)} - d, \theta_{(k-t+1)} + d]$ . See Figure 1 for a graphical representation of these intervals. A correct selection occurs when

$${}_d\text{CS}_t = \{A_{t_1} \subseteq s \text{ and } s \setminus A_{t_1} \subseteq A_{t_2}\},$$

where  $\setminus$  denotes the set difference operator  $B \setminus C = \{x : x \in B \text{ and } x \notin C\}$ . If a set  $s$  satisfies  ${}_d\text{CS}_t$ , then it is said to be a  $d$ -best set. The probability of selecting a  $d$ -best set is  $P({}_d\text{CS}_t)$ .

This selection goal is more complex. To illustrate, let  $t = 3$  and  $d = 0.5$ . Our  $s$  remains the same, because our former top three statistics are still the top three. So  $s = \{1, 3, 4\}$ . Referring to Table 2, we see that  $A_{t_1}$  would be the set of indices of the parameters in the interval  $(\theta_{(3)} + 0.5, \theta_{(5)}) = (3.00, 4.20)$ . So  $A_{t_1} = \{4\}$ . Then  $A_{t_2}$  would be the set of indices of the parameters in the interval  $[\theta_{(3)} - 0.5, \theta_{(3)} + 0.5] = [2.00, 3.00]$ . So  $A_{t_2} = \{1, 3, 5\}$ . Our result is that  $A_{t_1} = \{4\} \subseteq s$  and  $s \setminus A_{t_1} = \{1, 3\} \subseteq A_{t_2} = \{1, 3, 5\}$ , resulting in a correct selection.

This selection goal has different advantages and disadvantages than  $G$ -best selection. Unlike a  $G$ -best set, a  $d$ -best set could contain indices of populations that are not actually in the top  $t$ , and exclude some that are in the top  $t$ . The population with the highest parameter must be chosen for the set to be considered a correct selection though. Furthermore,  $d$ -best selection ensures that the populations deemed a correct selection are within  $d$  of the best parameters. The situation for using  $d$ -best selection would be when a selection of  $t$  populations is desired, and the difference of  $d$  units between parameters is unimportant. For example, selecting the absolute top ten best performing stocks might be virtually impossible (low  $P({}_d\text{CS}_t)$ ), but the ten best within \$0.50 might have a reasonable chance of success. Because fifty cents is negligible, the margin of error is acceptable.

**2.4. Calculation of  $G$ -best and  $d$ -best selection.** The formula for  $P(\text{CS}_{G,t})$  and  $P({}_d\text{CS}_t)$  is the same, but the calculations will differ based on the definitions of



$G$ -best and  $d$ -best sets. The formula is

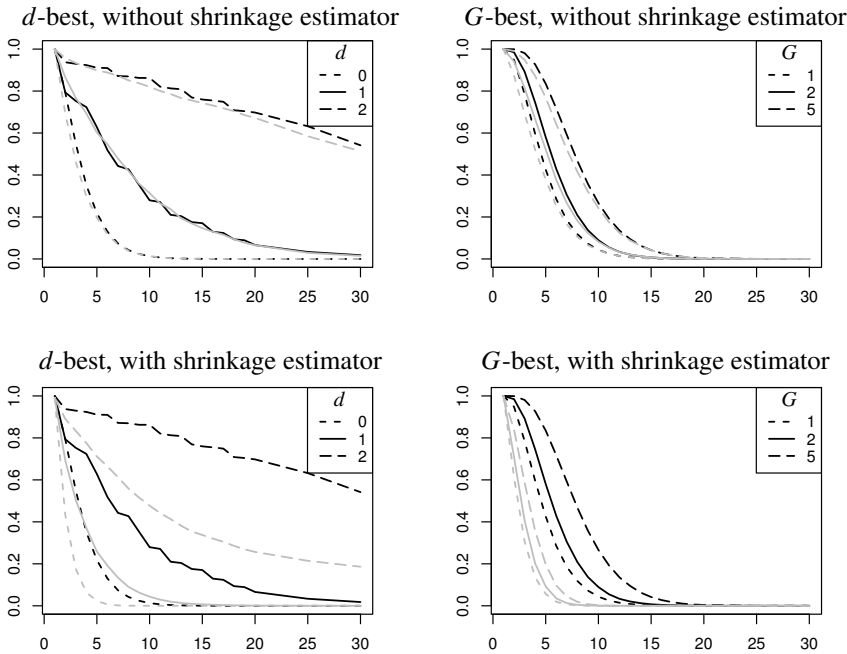
$$\sum_{g=1}^{|S|} \int_{-\infty}^{\infty} \prod_{j=k-t+1}^k [1 - F(y - \theta_{s_{g,j}})] d \left\{ \prod_{j=1}^{k-t} F(y - \theta_{\bar{s}_{g,j}}) \right\}. \tag{1}$$

We use  $S$  to denote the set of all  $G$ -best or  $d$ -best sets, or all the sets  $s$  that will result in a correct selection. Furthermore, let  $\theta_g$  be the  $g$ -th combination of ordered parameters. This will be a set of sets, each of which will contain the highest parameter  $\theta_{(k)}$  and be of size  $t$ . Then  $\theta_g = \{\theta_{s_{g,j}}, \theta_{\bar{s}_{g,j}}\}$ , where  $\theta_{s_{g,j}}$  denotes the combinations of parameters that satisfy the specific sets  $s_{g,j} \in S$  and  $\theta_{\bar{s}_{g,j}}$  contains the combinations of parameters that do *not* satisfy  $G$ -best or  $d$ -best sets. That is,  $\bar{s}_{g,j} \in \bar{S}$ .  $F(y - \theta_{s_{g,j}})$  is the continuous cumulative distribution function of the statistic  $y$ , adjusted to center around 0. See [Cui and Wilson 2008] for the derivation of (1).

These are extremely difficult integrals to integrate, analytically and even numerically. There are expansions that simplify the expression in order to make a numerical solution possible (specifically, via Gauss–Hermite quadrature [Cui and Wilson 2008]). To calculate the PCS for our datasets we use an R package that uses a parametric bootstrapping method.

**2.5. Performance of  $G$ -best and  $d$ -best selection.** A simulation study has been performed in [Cui and Wilson 2009] to assess the performance of both  $G$ -best and  $d$ -best selection. In this study, it was shown that for populations of a known parametric distribution the estimated PCS was accurate for both  $G$ -best and  $d$ -best selection when the distributional assumptions were met. Figure 2 shows one simulation of normal data with normal estimated PCS, done in [Cui and Wilson 2009]. Note that for this simulation  $n = 3$  and  $\sigma^2 = 3$ , which means the standard error is 1. Thus, when  $n > \sigma^2$ , the error decreases, and this was something that could be controlled. Another aspect of PCS that was studied in [Cui and Wilson 2009] was the use of shrinkage estimators. Shrinkage estimators are functions of the statistics designed to decrease the bias in the estimated PCS. The second row of graphs in Figure 2 uses a Stein-type shrinkage estimator and shows an increase in error compared to the PCS calculated with no shrinkage estimator. Cui and Wilson [2009] showed that this increase in error was characteristic for four different shrinkage estimators, which did not in general improve the bias. Thus, these shrinkage estimators were not recommended, and are not used in this paper. It was also shown that the PCS for  $G$ -best selection was still accurate when the normality assumption was violated, but the PCS for  $d$ -best selection had high error for high values of  $d$ .

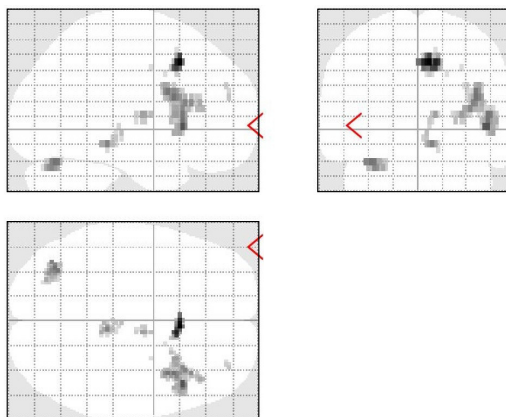
In the case that distributional assumptions are not satisfied, a nonparametric method may be used through bootstrapping if the sample size is large enough. In this case, a sufficiently narrow 95% confidence interval for the PCS could be found for large values of  $t$ . For example, one simulation consisted of two groups of 100 normal



**Figure 2.** Simulated normal data with normal estimated PCS, with  $t$  on the horizontal axes and PCS on the vertical. Note that  $n = 3$ ,  $\sigma^2 = 3$ , so  $SDE = 1$ . The black lines indicate the true PCS values and the gray lines indicate the estimated PCS in all graphs.

populations with variance 1 at a distance of three standard deviations away from one another. The 95% confidence interval for  $P(d=1 CS_{t=100})$  was 0.69–1.00. The 95% confidence interval for  $P(CS_{G=10,t=100})$  was 0.74–1.00 [Cui and Wilson 2009].

**2.6. The improvements for massive datasets.** The use of  $G$ -best and  $d$ -best selection is very practical with massive datasets. Ranking and selection methods to reduce the number of populations to study can be used along with multiple testing methods, but may even suit the needs of the researcher better than these methods. Furthermore,  $G$ -best and  $d$ -best selection are an improvement on previous definitions of PCS with respect to massive datasets. First of all, because  $G$ -best and  $d$ -best sets are defined in terms of index sets, they deal with the problem of having two equal parameters effectively. Also,  $d$ -best selection is especially useful for a dataset with high density, which is characteristic of massive datasets. The more the population parameters are approximately equal to other parameters, the more dense the data is. A researcher may not actually be interested in the absolute top  $t$  populations, but rather which populations will be most worthwhile to study. With  $d$ -best selection, the researcher can choose an interval around the true top



**Figure 3.** View of a brain 3D scan, showing the coordinate cross-sections containing a certain voxel (marked with  $<$ ) situated in the front left of the brain, a little less than halfway from the bottom.

parameters that is allowably close to find populations that may not be the best, but will be worth spending time on.

### 3. Application

**3.1. Introduction.** Although  $G$ -best and  $d$ -best selection are fully generalizable, in the literature to date they have only been applied to microarray data. To test and illustrate the applicability and usefulness of  $P(CS_{G,t})$  and  $P(dCS_t)$ , we will calculate the probability of correct selection in neuroimaging [Nichols and Hayasaka 2003] and econometrics [Romano and Wolf 2005] data. To calculate the probability of correct selection for these applications, we use the R package PCS, which can be found at [www.r-project.org](http://www.r-project.org).

**3.2. Neuroimaging.** First of all, we look at brain scans from a test on verbal fluency. Scientists conducted the study on five people, who both listened passively and said words aloud. They then studied whether areas of the brain were activated more in listening to or in generating words. To study the brain, they used 3D scans composed of *voxels*, which are three-dimensional pixels. Figure 3 shows the shaded voxels that represent activated areas of the brain common to all five subjects.

Nichols and Hayasaka [2003] took this study and measured each of the 55,027 voxels of the brain scans to see if any part of the brain was more active for word generation as opposed to passive listening. With this particular experiment, no voxels were found to be significant. To analyze these results with PCS, we used the program SPM8 [Friston et al. 2013] to find the possible clusters of voxels that might be significant in the conjunction of all five brains. We then calculated the probability that

$t$	1	2	3	4
$P(\text{CS}_{0,t}) = P({}_0\text{CS}_t)$	0.11	0.02	0.00	0.00
$P(\text{CS}_{2,t})$	0.33	0.11	0.03	0.01
$P(\text{CS}_{4,t})$	0.49	0.22	0.08	0.03
$P(\text{CS}_{6,t})$	0.60	0.33	0.15	0.06
$P({}_{0.5}\text{CS}_t)$	0.19	0.04	0.05	0.03
$P({}_1\text{CS}_t)$	0.24	0.19	0.35	0.29

**Table 3.** These low probabilities support Nichols' and Hayasaka's findings. For example, when  $t = 3$  and  $G = 4$ , the probability that the clusters selected actually are best is only .08.

one, two, three or four of these clusters may actually be the best clusters of voxels, or show the most brain activity. The calculation of PCS for this data supports Nichols' and Hayasaka's conclusion that there were no significant voxels (see Table 3).

The very low probabilities in Table 3 show that if one were to choose even one cluster as significant, it would not likely be the best cluster of voxels. Consider the probability of correct selection of  $t = 1$  for both  $G = 2$  and  $d = 0.5$ . To understand the meaning of the probability of a  $G$ -best selection, refer to Definition 2.1. For  $t = 1$  and  $G = 2$ , the probability of correct selection is .33. In the table, this is denoted by  $P(\text{CS}_{2,1}) = .33$ . If one chooses the top three clusters of voxels, there is only a .33 probability that the best cluster is among them. For  $d$ -best selection, refer to Definition 2.2. With  $t = 1$  and  $d = 0.5$ , the probability of correct selection is .19. That is,  $P({}_{0.5}\text{CS}_1) = .19$ . The probability that the top cluster of voxels is even within a margin of 0.5 of the top clusters is only .19. This complements the multiple testing result that none of the voxels are significantly different from any of the others. The highest probability found was for  $t = 1$  and  $G = 6$ . These parameters result in a probability of more than half. This supports Nichols' and Hayasaka's findings, but adds the information that we would have to choose seven clusters just to find one that stands out.

**3.3. Econometrics.** The economics data we chose comes from the Center for International Securities and Derivatives Market from January 1992 to March 2004. There are 105 hedge funds, and each fund has 147 recorded returns, one from each month in the time period. Instead of simply recording the return on investment for each month, the data records the amount the return is above or below a certain benchmark. In this case, the benchmark is the risk-free rate, i.e., the rate of return on an investment with zero risk. Romano and Wolf [2005] used stepwise multiple testing procedures to find the top ten absolute best performing funds. They defined best as the fund with the largest return in excess of the benchmark. We have calculated the probability that the ten funds that Romano and Wolf chose are

Excess	Fund
1.70	Libra Fund
1.41	Private Investment Fund
1.36	Aggressive Appreciation
1.27	Gamut Investments
1.26	Turnberry Capital
1.14	FBR Weston
1.11	Berkshire Partnership
1.09	Eagle Capital
1.07	York Capital
1.07	Cabelli International

**Table 4.** The ten highest-performing funds from January 1992 to March 2004, ranked by average return in excess of the risk-free rate.

actually the top ten using different parameters of  $G$ -best and  $d$ -best selection, but always choosing ten funds. The index set of the ten chosen funds is

$$s = \{31, 105, 16, 8, 25, 101, 38, 4, 82, 57\},$$

and the chosen funds according to [Romano and Wolf 2005] are shown in Table 4. Table 5 shows the probabilities of correct selection.

From these probabilities, one can see that it is not very certain that all ten funds chosen are truly the top ten funds. The probability  $P(CS_{0,10}) = P({}_0CS_{10}) = .13$  shows that these ten funds only have a 13% chance of being the top ten. Using  $G$ -best selection, we can be confident that these top ten statistics contain the top five funds, but that only accounts for half of the funds chosen. Furthermore, by looking at  $d$ -best selection, there is an .86 probability that the top ten funds found with the absolute statistic are within two ranks of the top ten funds. This is actually a very large margin. The reason the probabilities are not that certain is because there is a large amount of variability in this statistic. To address this issue, Romano and Wolf propose standardizing the statistic.

Romano and Wolf studentize the absolute statistic, that is, they used the usual  $t$ -statistic, which is calculated by dividing the statistic used above by the standard error. Romano and Wolf estimate variance using a sophisticated method involving a time-series bootstrap, whose code is unavailable. Thus, for this paper we simply divide the first statistic by the usual standard error (standard deviation divided by  $\sqrt{n}$ ) of each fund. The top ten funds Romano and Wolf chose using the  $t$ -statistics were a completely disjoint set from the nonstudentized set of best funds. The index set of the top ten funds found using the usual standard error is

$$s = \{61, 60, 102, 100, 23, 30, 18, 22, 63, 46\}.$$

Absolute Statistic		Studentized Statistic	
$P(\text{CS}_{G,t})$	$P({}_d\text{CS}_t)$	$P(\text{CS}_{G,t})$	$P({}_d\text{CS}_t)$
$P(\text{CS}_{0,10}) = 0.13$	$P({}_0\text{CS}_{10}) = 0.13$	$P(\text{CS}_{0,10}) = 0.71$	$P({}_0\text{CS}_{10}) = 0.71$
$P(\text{CS}_{1,9}) = 0.29$	$P({}_{0,5}\text{CS}_{10}) = 0.32$	$P(\text{CS}_{1,9}) = 0.98$	$P({}_{0,5}\text{CS}_{10}) = 0.71$
$P(\text{CS}_{2,8}) = 0.53$	$P({}_1\text{CS}_{10}) = 0.52$	$P(\text{CS}_{2,8}) = 1.00$	$P({}_1\text{CS}_{10}) = 0.95$
$P(\text{CS}_{3,7}) = 0.74$	$P({}_{1,5}\text{CS}_{10}) = 0.78$	$P(\text{CS}_{3,7}) = 1.00$	$P({}_{1,5}\text{CS}_{10}) = 0.95$
$P(\text{CS}_{4,6}) = 0.92$	$P({}_2\text{CS}_{10}) = 0.86$	$P(\text{CS}_{4,6}) = 1.00$	$P({}_2\text{CS}_{10}) = 0.97$
$P(\text{CS}_{5,5}) = 1.00$	$P({}_{2,5}\text{CS}_{10}) = 0.93$	$P(\text{CS}_{5,5}) = 1.00$	$P({}_{2,5}\text{CS}_{10}) = 1.00$
$P(\text{CS}_{6,4}) = 1.00$	$P({}_3\text{CS}_{10}) = 0.99$	$P(\text{CS}_{6,4}) = 1.00$	$P({}_3\text{CS}_{10}) = 1.00$
$P(\text{CS}_{7,3}) = 1.00$	$P({}_{3,5}\text{CS}_{10}) = 1.00$	$P(\text{CS}_{7,3}) = 1.00$	$P({}_{3,5}\text{CS}_{10}) = 1.00$
$P(\text{CS}_{8,2}) = 1.00$	$P({}_4\text{CS}_{10}) = 1.00$	$P(\text{CS}_{8,2}) = 1.00$	$P({}_4\text{CS}_{10}) = 1.00$
$P(\text{CS}_{9,1}) = 1.00$	$P({}_{4,5}\text{CS}_{10}) = 1.00$	$P(\text{CS}_{9,1}) = 1.00$	$P({}_{4,5}\text{CS}_{10}) = 1.00$

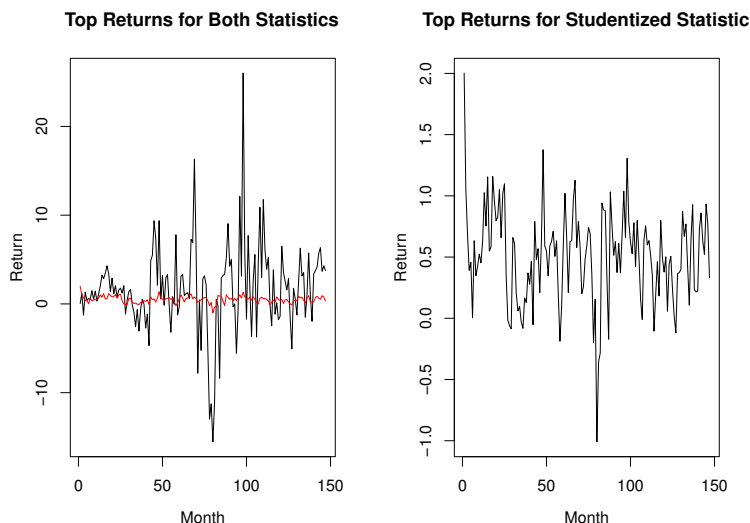
**Table 5.** The first two columns show the PCS for the absolute statistic. These probabilities are significantly lower than those from the studentized statistic, shown in the third and fourth columns. The difference is due to the studentized statistic accounting for the variability in the data.

Just as in [Romano and Wolf 2005], this is a completely different set of ten funds chosen as best. Table 5 shows the probability that the studentized statistic shows the true top ten studentized funds.

As one can see, the probability that these ten chosen funds are in actuality the best funds is significantly more than with the nonstudentized statistic. From these results, we can see that the studentized statistic, even using the usual standard error for each fund, is much more likely to identify the true best hedge funds according to the  $t$ -statistic.

Romano and Wolf show that the studentized statistic is a better measure of the performance of a fund because it takes into account the amount of risk involved. As one can see in Figure 4, the magnitude of the return of the top fund chosen according to the absolute statistic is much larger than that of the top fund chosen according to the studentized statistic. However, the second graph in Figure 4 shows that the return of the top fund according to the studentized statistic is in positive excess of the risk-free rate for the vast majority of the months recorded. The high probabilities found with PCS further support that the standardized statistic is superior to the absolute statistic. It is also important to note that the PCS of the studentized statistics may change with the more sophisticated estimate of variance. Still, even in this application, PCS provides useful information on both the absolute and studentized statistics.

**3.4. Results.** Applying PCS to these areas shows how useful ranking and selection methodology can be. The probability of correct selection has thus far been consistent



**Figure 4.** The graph on the left shows the return of the fund chosen as best according to the absolute statistics (black) and of the fund chosen as best according to the studentized statistic (red). The graph on the right shows the return of the top fund chosen using the studentized statistic alone.

with the latest multiple testing procedures, as it is in the neuroimaging application. However, PCS provides different information than a multiple hypothesis test. In each application we found a measure of how accurate our chosen best populations actually are. Instead of simply choosing the best statistics, one can have a better idea of how close they are to actually being best. If a neuroimaging scientist actually found a significant cluster of voxels, he or she would know how unlikely it is for that cluster to be best. With the econometrics application, investors can see the probability that the ten funds chosen either contain the actual top  $t$  funds, or that they are within a certain margin of the actual top ten funds. This is valuable information that can help drive the development of a new hypothesis.

#### 4. Conclusions

The probability of correct selection is a useful tool in statistics, and we have striven to illustrate this through both the theory behind  $G$ -best and  $d$ -best selection and its application to differing areas. The use of PCS deals with the problem of massive datasets by accommodating dense datasets that may have many parameters in common or close enough to study. Furthermore,  $G$ -best and  $d$ -best selection are useful tools for a researcher with limited resources. Instead of having a list of significant populations too large to adequately study, one can actually find the

populations that are most likely to be the best. Depending on the needs of the researcher,  $G$ -best or  $d$ -best selection may be more useful. Both selection goals were found to be consistent with previous claims of significance in the neuroimaging application, which supported their validity. In the econometrics application, PCS provided information on two separate statistics in the same study, which showed its adaptability. In both applications, PCS provided additional information that was not available through hypothesis testing. With PCS, we gain an insight into the quality of the populations chosen as best by seeing how likely it is that they truly are best. Clearly, PCS is a powerful tool.

For further research, we would like to find the variance estimator for a time-series regression bootstrap in order to find the PCS of the top ten funds actually chosen by Romano and Wolf. We would also like to apply PCS to mass spectrometry and other large  $k$  populations found in the literature.

### References

- [Bechhofer 1954] R. E. Bechhofer, “A single-sample multiple decision procedure for ranking means of normal populations with known variances”, *Ann. Math. Statistics* **25** (1954), 16–39. MR 15,638b Zbl 0055.13003
- [Cui and Wilson 2008] X. Cui and J. Wilson, “On the probability of correct selection for large  $k$  populations, with application to microarray data”, *Biom. J.* **50**:5 (2008), 870–883. MR 2542350
- [Cui and Wilson 2009] X. Cui and J. Wilson, “A simulation study on the probability of correct selection for large  $k$  populations”, *Comm. Statist. Simulation Comput.* **38**:6-7 (2009), 1244–1255. MR 2749857 Zbl 1167.62020
- [Friston et al. 2013] K. J. Friston, J. Ashburner, G. Flandin, J. Heather, A. Holmes, and J.-B. Poline, “Statistical parametric mapping”, 2013, available at <http://www.fil.ion.ucl.ac.uk/spm>.
- [Gupta 1956] S. S. Gupta, *On a decision rule for a problem in ranking means*, Ph.D. thesis, University of North Carolina, Chapel Hill, 1956. MR 2938890 Zbl 0073.35901
- [Nichols and Hayasaka 2003] T. Nichols and S. Hayasaka, “Controlling the familywise error rate in functional neuroimaging: a comparative review”, *Stat. Methods Med. Res.* **12**:5 (2003), 419–446. MR 2005445 Zbl 1121.62645
- [Romano and Wolf 2005] J. P. Romano and M. Wolf, “Stepwise multiple testing as formalized data snooping”, *Econometrica* **73**:4 (2005), 1237–1282. MR 2149247 Zbl 1153.62310

Received: 2011-08-29      Revised: 2013-07-12      Accepted: 2013-07-26

erin.c.irwin@biola.edu

*Mathematics and Computer Science Department,  
Biola University, 13800 Biola Avenue, La Mirada, CA 90639,  
United States*

jason.wilson@biola.edu

*Mathematics and Computer Science Department,  
Biola University, 13800 Biola Avenue, La Mirada, CA 90639,  
United States*



# On attractors and their basins

Alexander Arbieto and Davi Obata

(Communicated by Kenneth S. Berenhaut)

We prove that the map assigning to a given vector field the Lebesgue measure of the union of the basins of its attractors is lower semicontinuous in a residual subset of vector fields. Moreover, we prove that the Lebesgue measure of the union of the basins of attractors of a generic sectional axiom A vector field is total. For this, we also improve a result of Morales about sectional-hyperbolic sets. We also remark that homoclinic classes are topologically ergodic and that for a generic tame diffeomorphism, the union of the stable manifolds of the hyperbolic periodic orbits is dense in the manifold.

## 1. Introduction

One of the key notions in the theory of dynamical systems is that of attractors. By definition, an attractor captures the asymptotic information of a large set of orbits, called its basin, which always contains an open set. As an example, if an attractor is hyperbolic, then the asymptotic behavior of an orbit in its basin is governed by the dynamics of one orbit inside it (a shadowing property).

Moreover, that essentially every orbit is attracted by one attractor and that the set of attractors is finite (and possibly hyperbolic) implies that the dynamics of the system are nicely described by the attractors. For instance, this led Palis [2005] to conjecture that “there is a dense set  $D$  of dynamical systems such that any element of  $D$  has finitely many attractors whose union of basins of attraction has total probability”.

Mathematicians have made many efforts to understand attractors and their basins, not only for finite-dimensional dynamics, but also for PDEs (infinite-dimensional dynamical systems). See, for instance, [Constantin et al. 1985] or [Hale 2000].

On the other hand, to understand properties of the entire set of dynamical systems is a difficult task, and it is more reasonable to try to understand a large part of the set of dynamical systems. This reasoning leads to the theory of generic dynamics. Since

---

*MSC2010:* 37C10, 37C20.

*Keywords:* attractors, basin, sectional axiom A, sectional hyperbolic, basin of attraction.  
Partially supported by CNPq, FAPERJ and PRONEX/DS from Brazil.

the  $C^r$ -topology turns the space of diffeomorphisms (or vector fields) into a Baire space, it is natural to show that some properties holds for a residual subset of the space of dynamical systems, i.e., a countable intersection of open and dense subsets, since this will show the presence of this property for a dense subset and this property could be used to show another property in another residual subset. Indeed, the intersection of two residual subsets is also a residual subset. Usually, we say that a property holds for a generic system if it holds in a residual subset of dynamical systems.

The purpose of this article is to give some remarks about attractors, and their basins, of certain classes of dynamical systems, both diffeomorphisms and vector fields. These remarks are the results obtained by Obata [2010], guided by Arbieto, in his undergraduate monograph. We will state the results and refer the reader to the next section for the precise definitions of the more technical objects used in the statements.

Let  $M$  be a Riemannian closed manifold. We denote by  $\text{Diff}^1(M)$  the space of diffeomorphisms and by  $\mathfrak{X}^1(M)$  the space of vector fields, both endowed with the  $C^1$ -topology. We denote by  $m$  the Lebesgue measure and by  $d$  the geodesic distance, both induced by the Riemannian metric. If  $X \in \mathfrak{X}^1(M)$ , we denote by  $X_t$  the flow generated by  $X$ .

**Results for flows.** An attractor is an invariant compact subset  $\Lambda$  of  $M$  such that there exists a neighborhood  $U$  of  $\Lambda$  with

$$X_t(\bar{U}) \subset U \text{ for } t > 0 \quad \text{and} \quad \bigcap_{t \geq 0} X_t(U) = \Lambda.$$

The set  $U$  is called the *local basin* of  $\Lambda$  and  $B(\Lambda) := \bigcup_{t \leq 0} X_t(U)$  is the *basin* of  $\Lambda$ . We also define a set  $R$  to be a repeller if  $R$  is an attractor for  $-X$ .

Let  $X$  be a vector field, and denote by  $m(B(X))$  the Lebesgue measure of the union of the basins of the attractors of  $X$ . This generates a map  $\Phi : \mathfrak{X}^1(M) \rightarrow [0, +\infty]$ , defined as  $\Phi(X) := m(B(X))$  if there exists an attractor and  $\Phi(X) := 0$  if not.

**Theorem 1.** *There exists a residual subset  $\mathcal{R}$  such that  $\Phi|_{\mathcal{R}}$  is lower semicontinuous.*

The analogous statement holds for diffeomorphisms using the same proof.

Metzger and Morales [2008] extended the notion of axiom A vector fields for flows with singularities, called *sectional axiom A* vector fields. As an intermediate step to studying sectional axiom A vector fields, we have the following result:

**Theorem 2.** *There exists a residual subset  $\mathcal{R}$  such that if  $X$  is in  $\mathcal{R}$  and  $\Gamma = \Lambda_1 \cup \dots \cup \Lambda_k$ , with  $\Gamma \subset \Omega(X)$ , is a disjoint union of homogeneous sectional-hyperbolic sets for  $X$  or  $-X$ , and  $\Gamma$  is a proper subset of  $M$ , then  $m(\Gamma) = 0$ .*

We remark that it is well known that if  $M$  is a closed manifold which is a sectional-hyperbolic set for  $X$ , then  $X$  has no singularities and  $X$  is Anosov [Bautista and Morales 2011].

To prove this theorem, we extend a result of [Morales 2007]; see Theorem 13.

As a corollary, we obtain the following result, which improves Corollary D of [Alves et al. 2007] in two ways. We do not require that the vector field be  $C^{1+\varepsilon}$  or that the dimension of the manifold be 3. Indeed, in [Alves et al. 2007] it is proved that if a sectional axiom A vector field  $X$  over  $M^3$  is  $C^{1+\varepsilon}$ , then the Lebesgue measure of the union of the basins of its hyperbolic or sectional-hyperbolic attractors is total. We remark also that the union of the sets of  $C^{1+\varepsilon}$  vector fields, over any  $\varepsilon > 0$ , is a meager subset of vector fields.

**Theorem 3.** *Let  $X$  be a generic sectional axiom A vector field. Then either  $X$  is Anosov, or the Lebesgue measure of the nonwandering set of  $X$  is zero and the Lebesgue measure of the union of the basins of its attractors is total.*

A difficulty in proving this theorem is that it is not known whether the set of sectional axiom A vector fields (without cycles) is open. This is an interesting question. Even so, there are open sets of vector fields formed by sectional axiom A sets [Bautista and Morales 2011]. Moreover, [Morales and Pacifico 2003] shows that in dimension 3, generically, either a vector field has infinitely many sinks or sources or it is sectional axiom A. So, we obtain the following corollary:

**Corollary 4.** *If  $\dim(M) = 3$ , a generic vector field either has infinitely many sinks or sources, or the Lebesgue measure of the union of the basins of its attractors is total.*

**Results for diffeomorphisms.** Abdenur [2003] proved that attractors for generic diffeomorphisms are homoclinic classes. These classes are always transitive. However it can be proved that they have another property called topological ergodicity.<sup>1</sup>

**Proposition 5.** *Any homoclinic class of a periodic point  $p$ , with period  $k$ , of a diffeomorphism  $f$  is topologically ergodic. Moreover, for any two open sets  $U$  and  $V$ , the density of  $N(U, V) = \{i \geq 1 : f^i(U) \cap V \neq \emptyset\}$  is bounded by below  $1/k$ .*

Finally, the techniques used in the proof of the results above can be used to prove a folklore result. Since, as far as the authors know, it was never written, we include here a proof of this result:

**Proposition 6.** *If  $f$  is a  $C^1$ -generic tame diffeomorphism, then the union of the stable manifolds of the hyperbolic periodic orbits is dense in  $M$ .*

We observe that this result was proved in a more general setting (partially hyperbolic diffeomorphisms with one-dimensional central bundle) by Bonatti, Gan and Wen [Bonatti et al. 2007]. In particular, they obtain this corollary using stronger

<sup>1</sup>Recently Abdenur and Crovisier [2012] investigated the mixing property for isolated sets.

methods. However, the short proof given here only uses the connecting lemma. This is a particular case of Bonatti's conjecture; see [Bonatti et al. 2007].

**Conjecture 7.** *There exists a residual subset  $\mathcal{R} \subset \text{Diff}^1(M)$  such that for any  $f \in \mathcal{R}$ , the union of the stable manifolds of the hyperbolic periodic orbits is dense in  $M$ .*

This paper is organized as follows. In Section 2, we give precise definitions of terms used in the introduction. In Section 3, we prove Theorem 1. In Section 4, we prove Theorems 2 and 3 and also prove an extension of a theorem by Morales. In Section 5, we give a proof of Proposition 5. Finally, in Section 6, we give a proof of Proposition 6.

## 2. Preliminaries

In this section, we give precise definitions of terms used in the introduction and collect some useful results.

**2.1. Topology.** As remarked before, both  $\text{Diff}^1(M)$  and  $\mathfrak{X}^1(M)$  are Baire spaces. We will say that a property  $P$  is generic if it holds for a residual subset of these spaces. If the residual subset is fixed, we also say that an element of it is generic.

Let  $F(M)$  denote the space of compact subsets of  $M$ ; it is a metric space under the Hausdorff metric, given by

$$d_H(A, B) = \max\{d_A(B), d_B(A)\} \quad \text{for all } A, B \in F(M),$$

where  $d_A(B) = \max_{b \in B} \{\min_{a \in A} (d(a, b))\}$ .

Let  $(N, d)$  be a metric space. A map  $\varphi : N \rightarrow F(M)$  is *lower semicontinuous* at  $y \in N$  if  $y_n \rightarrow y$  implies  $d_H(\varphi(y_n), \varphi(y)) \rightarrow 0$ . Analogously, a map  $\varphi : N \rightarrow \mathbb{R}$  is lower semicontinuous at  $x_0 \in X$  if

$$\liminf_{x \rightarrow x_0} f(x) \geq f(x_0).$$

It is well known that if  $(N, d)$  is a Baire space, then the set of continuity points of a lower semicontinuous map, in either definition above, is a residual subset of its domain; see [Kelley 1955].

**2.2. Flows.** Let  $X \in \mathfrak{X}^1(M)$ . The orbit of a point  $p$  is the set  $\{X_t(p)\}_{t \in \mathbb{R}}$ . A *periodic orbit* of  $X$  is an orbit  $\{X_t(p) : t \in \mathbb{R}\}$  of a point  $p \in M$  satisfying  $X_T(p) = p$  for some minimal  $T > 0$ . A singularity  $\sigma$  is a zero of  $X$ . By a *closed orbit* we mean a periodic orbit or a singularity. The nonwandering set of  $X$  is the set  $\Omega(X)$  of points  $x$  such that for every neighborhood  $U$  of  $x$  and  $N > 0$ , there exists some  $T > N$  such that  $X_T(U) \cap U \neq \emptyset$ .

A subset  $\Lambda \subset M$  is *invariant* if  $X_t(\Lambda) = \Lambda$  for all  $t \in \mathbb{R}$ ; *transitive* if there exists  $p \in \Lambda$  such that its orbit is dense in  $\Lambda$ ; *isolated* if there exists a neighborhood  $U$

of  $\Lambda$  such that  $\bigcap_{t \in \mathbb{R}} X_t(U) = \Lambda$ ; and  $\Omega$ -isolated if there exists a neighborhood  $V$  of  $\Lambda$  such that  $\Omega(X) \cap V = \Lambda$ . We remark that any attractor is  $\Omega$ -isolated.

We say that a subset  $\Lambda \subset M$  is sectional-hyperbolic if every singularity in  $\Lambda$  is hyperbolic and it has a nontrivial partially hyperbolic splitting  $T_\Lambda M = E \oplus F$  such that  $E$  is uniformly contracting and  $F$  is sectionally expanding; i.e.,

$$\dim(E_x^c) \geq 2 \quad \text{and} \quad |\det(DX_t(x)/L_x)| \geq K^{-1}e^{\lambda t}$$

for all  $x \in \Lambda, t \geq 0$ , and  $L_x$  a two-dimensional subspace of  $E_x^c$ .

We say that  $\Lambda$  is *hyperbolic* if there is a continuous invariant tangent bundle decomposition

$$T_\Lambda M = \hat{E}_\Lambda^s \oplus \hat{E}_\Lambda^X \oplus \hat{E}_\Lambda^u,$$

and positive constants  $K, \lambda$ , where  $\hat{E}_\Lambda^X$  is the subbundle generated by  $X$  and

$$\|DX_t(x)/\hat{E}_x^s\| \leq Ke^{-\lambda t} \quad \text{and} \quad \|DX_{-t}(x)/\hat{E}_{X_t(x)}^u\| \leq Ke^{-\lambda t}$$

for all  $x \in \Lambda$  and  $t \geq 0$ .

A closed orbit is hyperbolic if it is a hyperbolic compact invariant set. A hyperbolic set is a basic set if it is isolated and transitive. Similar notions hold for diffeomorphisms.

Given an invariant splitting  $T_\Lambda M = E_\Lambda \oplus F_\Lambda$  over an invariant set  $\Lambda$  of a vector field  $X$ , we say that the subbundle  $E_\Lambda$  *dominates*  $F_\Lambda$  if there are positive constants  $K, \lambda$  such that

$$\|DX_t(x)/E_x\| \|DX_{-t}(x)/F_{X_t(x)}\| \leq Ke^{-\lambda t} \quad \text{for all } x \in \Lambda \text{ and } t \geq 0.$$

In such a case we say that  $T_\Lambda M = E_\Lambda \oplus F_\Lambda$  is a *dominated splitting*.

We say that  $\Lambda$  is *partially hyperbolic* if it has a dominated splitting  $T_\Lambda M = E_\Lambda^s \oplus E_\Lambda^c$  whose dominating subbundle  $E_\Lambda^s$  is *contracting*, that is,

$$\|DX_t(x)/E_x^s\| \leq Ke^{-\lambda t} \quad \text{for all } x \in \Lambda \text{ and } t \geq 0.$$

Moreover, we call the central subbundle  $E_\Lambda^c$  *sectionally expanding* if

$$\dim(E_x^c) \geq 2 \quad \text{and} \quad |\det(DX_t(x)/L_x)| \geq K^{-1}e^{\lambda t}$$

for all  $x \in \Lambda, t \geq 0$ , and  $L_x$  a two-dimensional subspace of  $E_x^c$ .

**Definition 8.** We say that a compact and invariant set  $\Lambda$  of  $X$  is *sectional-hyperbolic* if every singularity contained in  $\Lambda$  is hyperbolic and it has a nontrivial partially hyperbolic set with a sectionally expanding central subbundle.

Now, we recall the notion of sectional axiom A vector field, given in [Metzger and Morales 2008]; see also [Morales et al. 1999].

**Definition 9.** A vector field  $X$  is sectional axiom A if there is a finite disjoint decomposition

$$\Omega(X) = \Lambda_1 \cup \dots \cup \Lambda_k,$$

where each  $\Lambda_i$  is a hyperbolic basic set or a sectional-hyperbolic attractor up to time reversion.

**2.3. Diffeomorphisms.** If  $p$  is a hyperbolic periodic point, of period  $k$ , of a diffeomorphism  $f$ , then its stable manifold is the set

$$W^s(p) = \{y \in M : d(f^{kn}(y), p) \rightarrow 0 \text{ as } n \rightarrow \infty\}.$$

This set is in fact an immersed manifold. The stable manifold of the orbit of  $p$  is the union of the stable manifolds of  $f^i(p)$  for  $i = 0, \dots, k-1$ , and it is denoted by  $W^s(O(p))$ . Analogously, we define the unstable manifold of  $p$  and the orbit of  $p$ .

**Definition 10.** The homoclinic class of  $p$  is the set

$$H(p, f) = \overline{W^s(O(p)) \cap W^u(O(p))}.$$

A diffeomorphism is *tame* if its nonwandering set decomposes as a finite number of homoclinic classes and finitely many sinks or sources. Analogous definitions hold for vector fields.

Given two nonempty open sets  $U$  and  $V$ , we define the set of times that the orbit of  $U$  visits  $V$  as

$$N(U, V) = \{i \geq 1 : f^i(U) \cap V \neq \emptyset\}.$$

The following definition can be found in [Abdenur and Crovisier 2012]:

**Definition 11.** An invariant and compact subset  $\Lambda$  of  $f$  is topologically ergodic if for every two nonempty open sets  $U, V \subset \Lambda$ , we have

$$\limsup_{n \rightarrow \infty} \frac{\#N(U, V) \cap \{1, \dots, n\}}{n} > 0.$$

### 3. Proof of Theorem 1

First we observe that since attractors are isolated, there are at most countably many of them. Let  $\Lambda_1, \Lambda_2, \dots$  be the attractors of a generic vector field  $X$ . Denote by  $B(\Lambda_1), B(\Lambda_2), \dots$  its basins.

We select  $\Lambda_1, \dots, \Lambda_r$  such that

$$\sum_{i=1}^r m(B(\Lambda_i)) \geq m(B(X)) - \varepsilon.$$

There exist compact sets  $K_1, \dots, K_r$  such that  $\Lambda_i \subset K_i \subset B(\Lambda_i)$  for  $i = 1, \dots, r$  and such that

$$m(B(\Lambda_i) - K_i) < \frac{\varepsilon}{r}.$$

Now, we recall a result of Abdenur. Actually, he works with diffeomorphisms, but his proof holds for vector fields with the necessary adaptations. Also, he states his theorem for  $\Omega$ -isolated transitive sets, but we will only state it in the case of attractors, which is the context here.

**Theorem 12** [Abdenur 2003]. *There exists a residual subset  $\mathfrak{R} \subset \mathfrak{X}^1(M)$  such that if  $X \in \mathfrak{R}$  and  $\Lambda$  is an attractor of  $X$  with local basin  $U$  which does not reduce to a singularity, then there exists a neighborhood  $\mathfrak{U}$  of  $X$  such that for any  $Y \in \mathfrak{U} \cap \mathfrak{R}$ ,  $\Lambda(Y) = \bigcap_{t \geq 0} Y_t(U)$  is an attractor. Moreover, there exists a periodic orbit  $O(p)$  such that  $\Lambda(Y) = H(O(p), Y)$ .*

Thus, there are local basins  $U_i$  of  $\Lambda_i$  such that these local basins persist in a  $C^1$ -generic neighborhood of  $X$ . Since  $B(\Lambda_i) = \bigcup_{t \geq 0} X_{-t}(U_i)$  and  $K_i \subset B(\Lambda_i)$  is a compact set, there is  $T > 0$  such that

$$K_i \subset \bigcup_{t \in [0, T]} X_{-t}(U_i).$$

The set on the right is open. So, if  $Y$  is  $C^1$ -close to  $X$ , we obtain that

$$K_i \subset \bigcup_{t \in [0, T]} Y_{-t}(U_i).$$

Thus, if  $Y$  is generic and  $C^1$ -close to  $X$ , we have  $m(B(\Lambda_i(Y))) \geq m(K_i)$ .

Hence,

$$m(B(Y)) \geq \sum_{i=1}^r m(K_i) \geq m(B(X)) - \varepsilon.$$

This proves lower semicontinuity.

#### 4. Proof of Theorems 2 and 3

Let  $\Lambda$  be a sectional-hyperbolic set for  $X$ . We recall that its strong stable manifold is the set

$$W^{ss}(x) = \left\{ y \in M : \lim_{t \rightarrow \infty} d(X_t(x), X_t(y)) = 0 \right\}.$$

Its local strong stable manifold is an  $\varepsilon$ -ball  $W_\varepsilon^{ss}(x)$  in  $W^{ss}(x)$  centered at  $x$  for some  $\varepsilon > 0$ .

Given  $A \subset M$ , we define  $\alpha(A)$  as the set of points  $y = \lim_{n \rightarrow \infty} X_{t_n}(z_n)$  for some sequences  $t_n \rightarrow -\infty$  and  $z_n \in A$ . We say that a sectional-hyperbolic set is homogeneous if the splitting  $E^s \oplus E^c$  given by the definition is such that  $\dim E^s$  is constant.

The following result improves the main theorem in [Morales 2007] since we do not require transitivity.

**Theorem 13.** *Let  $\Lambda \subset \Omega(X)$  be a homogeneous sectional-hyperbolic set for  $X$ . Denote by  $R$  the union of the hyperbolic repellers contained in  $\Lambda$ . Then  $\Lambda - R$  does not contain any local strong stable manifold.*

*Proof.* By hypothesis, the map  $x \in \Lambda \mapsto W_\varepsilon^{\text{ss}}(x)$  is continuous if  $\varepsilon > 0$  is small, but fixed. Assume that  $\Lambda - R$  contains some  $W_\varepsilon^{\text{ss}}(x)$ . Let  $\delta < \varepsilon$  and take  $H = \alpha(W_\delta^{\text{ss}}(x)) \subset \Lambda - R$ , which is compact and invariant. Observe also that the set  $\Lambda - R$  is compact and invariant, since  $\Lambda \subset \Omega(X)$ .

If  $H$  has a singularity  $\sigma$  then, by definition,  $\sigma = \lim X_{t_n}(z_n)$  for some sequences  $t_n \rightarrow -\infty$  and  $z_n \in W_\delta^{\text{ss}}(x)$ . Moreover,  $W_\delta^{\text{ss}}(X_{t_n}(z_n)) \subset \Lambda$  for any natural number  $n$ . Taking the limit as  $n \rightarrow \infty$ , we obtain that  $W_\delta^{\text{ss}}(\sigma) \subset \Lambda$ .

However, by [Bautista and Morales 2011], since  $\Lambda$  is sectional-hyperbolic, we have that  $\Lambda \cap W^{\text{ss}}(\sigma) = \{\sigma\}$ , and this is a contradiction.

If  $H$  does not have a singularity, then by the hyperbolic lemma [Bautista and Morales 2011],  $H$  is a hyperbolic set. Now, let  $y$  be a cluster point of  $X_{t_n}(x)$ , with  $t_n \rightarrow -\infty$ . We will show that  $W^{\text{ss}}(y) \subset H$ . Indeed, let  $z \in W^{\text{ss}}(y)$  and let  $\varepsilon > 0$  be small enough. There exists  $T > 0$  such that

$$d(X_T(z), X_T(y)) < \varepsilon.$$

Also, there exists  $n_0$  such that for any  $n \geq n_0$ , we have

$$d(X_t(y), X_{t_n+T}(x)) < \varepsilon.$$

Finally, for any  $n$  large, there exists  $z_n \in W_\delta^{\text{ss}}(x)$  such that

$$d(X_{t_n+T}(x), X_{t_n+T}(z_n)) < \varepsilon.$$

This implies that if  $n$  is large enough then

$$d(X_T(z), X_{t_n+T}(z_n)) < \varepsilon.$$

In particular, we can assume that  $t_n + T \rightarrow -\infty$ . Thus,  $X_T(z) \in H$ , and by invariance,  $z \in H$ . Thus  $H$  is a repeller inside  $\Lambda - R$ , a contradiction.  $\square$

**Remark 14.** We could remove the homogeneity assumption. Indeed, the sets  $\{x \in \Lambda : \dim(E^s(x)) = i\}$  for  $1 \leq i \leq d - 1$  are compact. Hence, we could use the argument restricting ourselves to each of these sets.

Now, we observe that  $X_1$  is a partially hyperbolic diffeomorphism over  $\Lambda$  since the dominated splitting  $T_\Lambda M = E^s \oplus E^c$  has a contracting subbundle  $E^s$ . A strong stable disk of  $X_1$  is a disk which is tangent to the subbundle  $E^s$  over  $\Lambda$ . Obviously, a strong stable disk of  $X_1$  is a local strong stable manifold for some point  $x \in \Lambda$ . However, the following result was proved in [Alves et al. 2007, Theorem 2.2]:



**Theorem 15.** *Let  $f : M \rightarrow M$  be a  $C^2$  diffeomorphism and  $\Lambda \subset M$  a partially hyperbolic set with positive volume. Then  $\Lambda$  contains a strong stable disk.*

Together with Theorem 13, we obtain the following:

**Corollary 16.** *Let  $\Lambda$  be a proper subset of  $M$ . If  $\Lambda$  is a homogeneous sectional-hyperbolic set of a  $C^2$  vector field  $X$  and  $\Lambda \subset \Omega(X)$ , then  $m(\Lambda) = 0$ .*

*Proof.* First we remark that there are only countably many repellers in  $\Lambda$ , since they are isolated. Moreover, by [Bowen 1975], the measure of any hyperbolic repeller (or attractor) is zero if  $X$  is  $C^2$ .

On the other hand, if  $R$  denotes the union of the hyperbolic repellers of  $\Lambda$  and  $m(\Lambda - R) > 0$ , then by Theorem 15, there exists a strong stable disk on  $\Lambda - R$ , and this contradicts Theorem 13.  $\square$

For any open set  $U$ , let  $\Lambda_Y(U) = \bigcap_{t \in \mathbb{R}} Y_t(\bar{U})$ . These sets have an upper semi-continuity property:  $\limsup \Lambda_{X_n}(U) \subset \Lambda_X(U)$ . Indeed, let  $x \in \limsup \Lambda_{X_n}(U)$ . So, there exists  $x_n \in \Lambda_{X_n}(U)$  such that  $x_n \rightarrow x$ . Fix  $t \in \mathbb{R}$ . We have  $(X_n)_t(x_n) \in \bar{U}$ . Thus,  $X_t(x) \in \bar{U}$ . Since this holds for every  $t \in \mathbb{R}$ , this implies that  $x \in \Lambda_X(U)$ .

Now, let  $\{U_k\}$  be a countable basis of the topology and  $\{O_k\}$  the set of finite unions of the  $U_k$ . For every  $n, k \in \mathbb{N}$ , we define  $\mathcal{U}_{n,k}$  as the set of vector fields  $Y$  such that  $m(\Lambda_Y(O_k)) < 1/n$ .

**Lemma 17.**  *$\mathcal{U}_{n,k}$  is an open set.*

*Proof.* Let  $Y \in \mathcal{U}_{n,k}$ , and suppose that  $m(\Lambda_Y(O_k)) = 1/n - \varepsilon$ . There exists  $T$  large enough that

$$m\left(\bigcap_{t=-T}^T Y_t(\bar{O}_k)\right) < m(\Lambda_Y(O_k)) + \frac{\varepsilon}{2}.$$

Let  $W$  be a neighborhood of  $\bigcap_{t=-T}^T Y_t(\bar{O}_k)$  such that

$$m(W) < m\left(\bigcap_{t=-T}^T Y_t(\bar{O}_k)\right) + \frac{\varepsilon}{2}.$$

If  $Z$  is close enough to  $Y$ , we have that  $\bigcap_{t=-T}^T Z_t(\bar{O}_k) \subset W$ . Thus

$$\begin{aligned} m(\Lambda_Z(O_k)) &\leq m\left(\bigcap_{t=-T}^T Z_t(\bar{O}_k)\right) < m\left(\bigcap_{t=-T}^T Y_t(\bar{O}_k)\right) + \frac{\varepsilon}{2} \\ &\leq m(\Lambda_Y(O_k)) + \varepsilon = \frac{1}{n}. \end{aligned} \quad \square$$

Now, we prove Theorem 2.

*Proof of Theorem 2.* By the previous lemma,  $\mathcal{O}_{n,k}$  is an open set. Now, we define  $\mathcal{N}_{n,k} = \mathfrak{X}^1(M) - \overline{\mathcal{O}_{n,k}}$ . Consider the residual subset

$$\mathcal{R} = \bigcap_n \bigcap_k (\mathcal{O}_{n,k} \cup \mathcal{N}_{n,k}).$$

Let  $X \in \mathcal{R}$  and let  $\Gamma = \Lambda_1 \cup \dots \cup \Lambda_k$ , as in the statement of Theorem 2. Suppose that  $\Lambda_i$  is a homogeneous sectional-hyperbolic set for  $X$ . Since  $\Lambda_i$  is invariant, there exists  $k(i)$  such that  $\Lambda_i \subset \Lambda_X(\mathcal{O}_{k(i)})$  and  $\Lambda_X(\mathcal{O}_{k(i)})$  is a homogeneous sectional-hyperbolic set. A similar argument holds when  $\Lambda_i$  is a homogeneous sectional-hyperbolic set for  $-X$ .

Now, suppose that  $m(\Lambda_X(\mathcal{O}_{k(i)})) > 0$  for some  $i$ . Thus, there exists  $n$  such that  $m(\Lambda_X(\mathcal{O}_{k(i)})) \geq 1/n$ . So,  $X \in \mathcal{N}_{n,k(i)}$ . Since  $\mathcal{N}_{n,k(i)}$  is an open set, there exists a neighborhood  $\mathcal{V}$  of  $X$  such that  $m(\Lambda_Y(\mathcal{O}_{k(i)})) \geq 1/n$  for every  $Y \in \mathcal{V}$ .

Using the semicontinuity property, mentioned above, and the sectional hyperbolicity of  $\Lambda_X(\mathcal{O}_{k(i)})$ , we can assume, shrinking  $\mathcal{V}$  if necessary, that  $\Lambda_Y(\mathcal{O}_{k(i)})$  is a homogeneous sectional-hyperbolic set for every  $Y \in \mathcal{V}$ .

Now, we can choose a  $C^2$  vector field  $Y \in \mathcal{V}$  and by Corollary 16, we have that  $m(\Lambda_Y(\mathcal{O}_k)) = 0$ , a contradiction. □

*Proof of Theorem 3.* The arguments given above show that there exists a residual subset  $\mathcal{S}$  such that if  $X \in \mathcal{S}$  and  $\Lambda$  is a proper saddle-type isolated transitive sectional-hyperbolic set, then  $m(\Lambda) = 0$ .

Indeed, let  $U$  be an open set, and define  $\mathcal{U}(U)$  as the (open) set formed by vector fields  $Y$  such that  $\Lambda_Y(U)$  is hyperbolic of saddle type. Let  $\mathcal{U}_n(U) = \{Y \in \mathcal{U}(U) : m(B(\Lambda_Y(U))) < 1/n\}$ . Using the same argument as in the proof of Lemma 17, we obtain that  $\mathcal{U}_n(U)$  is an open set.

Moreover, if  $Y \in \mathcal{U}(U)$  is  $C^2$ , we have that  $m(B(\Lambda_Y(U))) = 0$  [Bowen 1975, p. 68]. So,  $\mathcal{U}_n(U)$  is dense in  $\mathcal{U}(U)$ .

Defining  $\mathcal{O}_k$  as above, we set  $\mathcal{S} = \bigcap_{k,n} \mathcal{U}_n(\mathcal{O}_k)$ .

Let  $X \in \mathcal{R} \cap \mathcal{S}$  be a sectional axiom A vector field. By definition, we have a spectral decomposition  $\Omega(X) = \Lambda_1 \cup \dots \cup \Lambda_k$ , formed by sectional-hyperbolic attractors, repellers and basic saddle-type hyperbolic sets. Moreover, since these sets have a dense orbit, they are homogeneous.

If  $m(\Omega(X)) > 0$ , then there exists  $1 \leq i \leq k$  such that  $m(\Lambda_i) > 0$ . By the previous argument and Theorem 2, we have that  $\Lambda_i = M$ . If  $\Lambda_i$  is a saddle-type hyperbolic set, then  $X$  is Anosov. If  $\Lambda_i$  is a sectional-hyperbolic attractor then it cannot have any singularity. Indeed, if  $\sigma$  is a singularity, we must have  $W^{ss}(\sigma) \cap \Lambda_i = \{\sigma\}$ , but if  $\Lambda_i = M$  this cannot be true. Hence, by the hyperbolic lemma,  $M$  would be hyperbolic again and  $X$  would be Anosov. If  $\Lambda_i$  is a sectional-hyperbolic attractor for  $-X$ , the same holds.

So, assuming that  $X$  is not Anosov,  $m(\Omega(X)) = 0$ . Using Lemma 2.2 of [Shub 1978], we have that

$$M = W^s(\Lambda_1) \cup \dots \cup W^s(\Lambda_k).$$

Since  $X \in \mathcal{R}$ , if  $\Lambda_i$  is a repeller then  $W^s(\Lambda_i) = \Lambda_i$  and  $m(\Lambda_i) = 0$ . Since  $X \in \mathcal{S}$ , if  $\Lambda_i$  is a hyperbolic basic set then  $m(W^s(\Lambda_i)) = 0$ . Thus, the measure of the union of the basins of the attractors is total.  $\square$

### 5. Proof of Proposition 5

In the following, we will work with the topology relative to the homoclinic class. First, we will show that any homoclinic class is topologically ergodic.<sup>2</sup>

Let  $H(p, f)$  be a homoclinic class. Denote by  $k$  the period of  $p$ . We recall that the local stable manifold of  $p$  is the set  $W_\varepsilon^s(p) = \{y \in M : d(f^n(y), f^n(p)) \leq \varepsilon\}$ .

Fix two nonempty open subsets  $U$  and  $V$  of  $H(p, f)$ . Since the stable manifold of its orbit is dense, there exist  $\varepsilon > 0$  and  $N > 0$  such that

$$f^{-N}(W_\varepsilon^s(p)) \cap U \neq \emptyset.$$

In particular, there exists a disk  $D \subset f^N(U)$  transversal to  $W_\varepsilon^s(p)$ . Moreover, since  $W^u(O(p))$  is dense, there exists  $K > 0$  such that  $f^K(W_\varepsilon^u(p)) \cap V \neq \emptyset$ . Using the  $\lambda$ -lemma [Palis and de Melo 1982], there exist  $m_0$  and  $0 \leq i < k$  such that for every  $m \geq m_0$ , we have that  $f^{km+i}(D) \cap V \neq \emptyset$ . Let  $l \in \mathbb{N}$  such that  $A = lk + (N + i)$  is the largest integer less than or equal to  $n$ ; in particular,  $n < A + k$ . By the previous remark,  $\#N(U, V) \cap \{1, \dots, n\} \geq l - m_0$ . So,

$$\limsup_{n \rightarrow \infty} \frac{\#N(U, V) \cap \{1, \dots, n\}}{n} \geq \limsup_{l \rightarrow \infty} \frac{l - m_0}{(l + 1)k + N + i} = \frac{1}{k}.$$

This shows that the homoclinic class is topologically ergodic.

### 6. Proof of Proposition 6

We recall that an invariant and compact subset  $A \subset M$  is called Lyapunov stable if given  $U$ , an open neighborhood of  $A$ , there exists another neighborhood  $V$  of  $A$  such that  $f^n(V) \subset U$  for every  $n \in \mathbb{N}$ .

**Lemma 18** [Carballo et al. 2003, Lemma 3.4]. *If  $f$  is a  $C^1$ -generic diffeomorphism, then  $\overline{W^u(O(p))}$  is Lyapunov stable for  $f$ .*

Another source of Lyapunov stable sets is the following, which is [Morales and Pacifico 2002, Theorem A]:

**Theorem 19.** *There exists a residual subset  $R^* \subset \text{Diff}^1(M)$  such that if  $g \in R^*$ , then the set  $S = \{x \in M : \omega(x) \text{ is Lyapunov stable}\}$  is a residual subset of  $M$ .*

<sup>2</sup>We want to thank Professor Abdenur for pointing out this short argument to us.

We also recall Hayashi’s connecting lemma [1997], one of the most useful techniques in the  $C^1$ -generic theory of dynamical systems. The formulation that we give here is taken from [Wen and Xia 2000].

**Theorem 20** (connecting lemma). *Let  $f \in \text{Diff}^1(M)$ , and let  $z$  be a nonperiodic point of  $f$ . Given a neighborhood  $\mathcal{U}$  of  $f$ , there exist  $\rho > 1$ ,  $L \in \mathbb{N}$  and  $\delta_0 > 0$  with the following property. Let  $0 < \delta < \delta_0$  and*

$$p, q \notin \Delta(\delta) := \bigcup_{n=1}^L (f^{-n}(B(z, \delta))).$$

*If there exist  $a > L$  such that  $f^a(p) \in B(z, \delta/\rho)$  and  $b \geq 0$  such that  $f^{-b}(q) \in B(z, \delta/\rho)$ , then there exists  $g \in \mathcal{U}$  such that  $q$  is a future  $g$ -iterate of  $p$  and  $g \equiv f$  outside  $\Delta(\delta)$ .*

We remark that the method used in the proof of Theorem 1 could be used to prove the topological semicontinuity of the basins of generic attractors. However, in the  $C^1$ -topology a stronger property can be obtained, which, together with the continuity given by the stable manifold theorem, quickly implies this semicontinuity in this topology.

**Proposition 21.**  *$C^1$ -generically, if a diffeomorphism has an attractor, then there exists a periodic point inside the attractor such that its stable manifold is dense in the basin of the attractor.*

*Proof.* Let  $U$  be an open set. We define the set

$$\mathcal{U}(U, m) = \left\{ f \in \text{Diff}^1(M) : \exists p \in \bigcap_{n \geq 0} f^n(U) \cap \text{Per}_h(f) \text{ with } W^s(p, f) \text{ } 1/m\text{-dense in } U \right\}.$$

If  $f \in \mathcal{U}(U, m)$ , then it has a hyperbolic periodic point in  $U$  such that its stable manifold is  $1/m$ -dense in  $U$ . Since this point is hyperbolic, there exists  $V$ , a  $C^1$ -neighborhood of  $f$  such that if  $g \in V$  then  $p(g) \in U$ . Take  $y \in U$  and  $B = B(y, 1/m)$ , so for  $f$ , we have that  $W^s(p, f) \cap B \neq \emptyset$ . By the stable manifold theorem, we have that  $W^s(p(g)) \cap B \neq \emptyset$ , so  $W^s(p(g))$  is  $1/m$ -dense in  $U$  and  $g \in \mathcal{U}(U, m)$ . This proves that the set  $\mathcal{U}(U, m)$  is open.

Let  $\{U_k\}$  be a countable basis of open sets of  $M$ , and let  $\{O_n\}$  be the set of all possible unions of the elements  $U_k$ . Define

$$A(O_n, m) = \mathcal{U}(O_n, m) \cup \overline{\mathcal{U}(O_n, m)^c}.$$

Now, by the previous remark, and by construction, this set is open and dense in  $\text{Diff}^1(M)$ . So  $R_1 = \bigcap_{n,m} A(O_n, m)$  is a residual subset. Let  $R_2$  be the residual subset given in [Abdenur 2003].

Let  $R = R_1 \cap R_2$ . If  $f \in R$  and  $\Lambda$  is an attractor of  $f$ , then there exists  $p \in \text{Per}_h(f)$  such that  $\Lambda = H(p, f)$ . Fix  $n, m$  such that  $O_n$  is a local basin of  $\Lambda$ . Now, we must prove that  $f \in \mathcal{U}(O_n, m)$ . Suppose that  $f \in \overline{\mathcal{U}(O_n, m)^c}$ . Since this set is open, there is  $W \subset \overline{\mathcal{U}(O_n, m)^c}$ , a small open  $C^1$ -neighborhood of  $f$ . The next step is to prove that we can find  $g \in W$  such that  $g \in \mathcal{U}(O_n, m)$ , which will be a contradiction. To prove this we will use the  $C^1$ -connecting lemma, and we will also need the following lemmas. From now on we will fix  $f$  and  $W$  as above.

**Lemma 22.** *The function  $\Phi(g) = \overline{W^s(p(g), g)}$  for  $g \in W$  is continuous in a residual subset of  $W$ .*

*Proof.* The map  $\Phi$  is lower semicontinuous in  $W$  by the stable manifold theorem. Then, it is continuous in a residual subset  $W^* \subset W$ . □

Thus we have that the map  $\Phi$  is continuous in  $W^* \cap R$ . Now, since  $f \in \overline{\mathcal{U}(O_n, m)^c}$ , there exists an  $x \in O_n$  such that  $B(x, 1/m) \cap \Phi(f) = \emptyset$ .

This, together with Theorem 19, implies the following corollary:

**Corollary 23.** *There exists a residual subset  $R_W \subset W$  such that if  $g \in R_W$ , then there exists a residual subset  $P \subset O_m$  such that if  $x \in P$  then  $\omega(x) = \Lambda(g)$ .*

*Proof.* Let  $R^*$  and  $S$  be given by Theorem 19. Define  $R_W := R \cap R^* \cap W$  and  $P = O_n \cap S$ . Hence, if  $x \in P$ , then  $x \in O_n$  and  $\omega(x) \subset \Lambda$ . However, since  $x \in S$  as well, we know that  $\omega(x)$  is Lyapunov stable. By the previous remark, since  $\Lambda$  is transitive, we have that  $\Lambda \subset \omega(x)$ . Thus  $\omega(x) = \Lambda$ . □

Now, we study the consequences of the continuity of  $\Phi$ .

**Lemma 24.** *If  $\Phi$  is continuous in  $g \in R_W$  and  $S$  is the set given by Theorem 19, then  $O_n \cap S \subset \Phi(g)$ .*

*Proof.* If the lemma does not hold, then there exists  $x \in (O_n \cap S) - \Phi(g)$ . Let  $U$  be a neighborhood of  $\Phi(g)$  such that  $x \notin U$ . By continuity there exists a neighborhood  $\mathcal{V}$  of  $g$  such that if  $h \in \mathcal{V}$  then  $\Phi(h) \subset U$ .

Since  $x \in O_n \cap S$ , we have  $\omega(x) = \Lambda(g)$ . Thus, there exists a sequence  $(l_n) \subset \mathbb{N}$  such that  $g^{l_n}(x) \rightarrow p(g)$ . By the Hartman–Grobman theorem [Palis and de Melo 1982], there exists another sequence  $(t_n) \subset \mathbb{N}$  such that

$$g^{t_n}(x) \rightarrow q \in W_\varepsilon^s(p(g), g) - \{p(g)\}.$$

Let  $\rho > 1$ ,  $L \in \mathbb{N}$  and  $\delta_0 > 0$ , as given by the  $C^1$ -connecting lemma applied to  $q$  and  $\mathcal{U}$ . Choose  $\delta$  with  $0 < \delta < \delta_0$ , and let  $V$  be a neighborhood of the orbit of  $p(g)$  such that

$$p(g), x \notin \Delta(\delta) = \bigcup_{n=1}^L (g^{-n}(B(q, \delta)))$$

and

$$\bigcup_{n=1}^L (g^{-n}(B(q, \delta))) \cap V = \emptyset.$$

Pick  $y \in B(q, \delta/\rho) \cap W^s(p(g), g)$  such that, defining  $z = g^k(y)$ , we have  $z \in (W^s(p(g), g) - \{p(g)\}) \cap V$ . By definition, we have that  $g^{-k}(z) = y \in B(q, \delta/\rho)$ . Using that  $g^{t_n}(x) \rightarrow q$ , we obtain some  $n_0 > L$  such that

$$g^{t_{n_0}}(x) \in B(q, \delta/\rho).$$

Applying the  $C^1$ -connecting lemma, we obtain  $h \in \mathcal{V}$  such that  $h = g$  outside of  $\Delta(\delta)$  and  $x$  belongs to the  $h$ -negative orbit of  $z$ . However, since  $z \in (W_g^s(p(g)) - \{p(g)\}) \cap V$ , we obtain that the  $h$ -positive orbit of  $z$  belongs to  $V$ .

Thus

$$z \in W_h^s(p(h)) \quad \text{and thus} \quad x \in W_h^s(p(h)).$$

This leads to a contradiction, since  $h \in \mathcal{V}$  and  $x \notin U$ . □

By the previous lemma, since  $f \in W$ , there is  $g \in R_W$  such that  $\Phi$  is continuous in  $g$ . So  $O_n \cap S \subset \Phi(g)$ , and there exists  $y \in B(x, 1/m) \cap S$ . Then  $y \in \Phi(g)$ , which is a contradiction since  $f \in W \subset \overline{\mathcal{U}(O_n, m)}^c$ . Then  $f \in \mathcal{U}(O_n, m)$ , which proves the proposition. □

Now, to prove Proposition 6, it is enough to combine Proposition 21 with:

**Theorem 25** [Carballo and Morales 2003]. *If  $f$  is a  $C^1$ -generic tame diffeomorphism then the union of the basins of its attractors is an open and dense subset of  $M$ .*

## References

- [Abdenur 2003] F. Abdenur, “Attractors of generic diffeomorphisms are persistent”, *Nonlinearity* **16**:1 (2003), 301–311. MR 2003k:37040 Zbl 1023.37007
- [Abdenur and Crovisier 2012] F. Abdenur and S. Crovisier, “Transitivity and topological mixing for  $C^1$  diffeomorphisms”, pp. 1–16 in *Essays in mathematics and its applications*, edited by P. M. Pardalos and T. M. Rassias, Springer, Heidelberg, 2012. MR 2975581
- [Alves et al. 2007] J. F. Alves, V. Araújo, M. J. Pacifico, and V. Pinheiro, “On the volume of singular-hyperbolic sets”, *Dyn. Syst.* **22**:3 (2007), 249–267. MR 2008k:37069 Zbl 1226.37014
- [Bautista and Morales 2011] S. Bautista and C. A. Morales, “Lectures on sectional-Anosov flows”, IMPA preprint series D 86/2011, Instituto Nacional de Matemática Pura e Aplicada, Rio de Janeiro, 2011, available at [http://preprint.impa.br/Shadows/SERIE\\_D/2011/86.html](http://preprint.impa.br/Shadows/SERIE_D/2011/86.html).
- [Bonatti et al. 2007] C. Bonatti, S. Gan, and L. Wen, “On the existence of non-trivial homoclinic classes”, *Ergodic Theory Dynam. Systems* **27**:5 (2007), 1473–1508. MR 2009d:37036 Zbl 1128.37021
- [Bowen 1975] R. Bowen, *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*, Lecture Notes in Math. **470**, Springer, Berlin, 1975. MR 56 #1364 Zbl 0308.28010

- [Carballo and Morales 2003] C. M. Carballo and C. A. Morales, “Homoclinic classes and finitude of attractors for vector fields on  $n$ -manifolds”, *Bull. London Math. Soc.* **35**:1 (2003), 85–91. MR 2003h:37021 Zbl 1035.37007
- [Carballo et al. 2003] C. M. Carballo, C. A. Morales, and M. J. Pacifico, “Homoclinic classes for generic  $C^1$  vector fields”, *Ergodic Theory Dynam. Systems* **23**:2 (2003), 403–415. MR 2004e:37031 Zbl 1047.37009
- [Constantin et al. 1985] P. Constantin, C. Foias, and R. Temam, *Attractors representing turbulent flows*, Mem. Amer. Math. Soc. **53**:314, American Mathematical Society, Providence, RI, 1985. MR 86m:35137 Zbl 0567.35070
- [Hale 2000] J. K. Hale, “Dissipation and attractors”, pp. 622–637 in *International Conference on Differential Equations* (Berlin, 1999), vol. 1, edited by B. Fiedler et al., World Scientific, River Edge, NJ, 2000. MR 1870207 Zbl 0971.37037
- [Hayashi 1997] S. Hayashi, “Connecting invariant manifolds and the solution of the  $C^1$  stability and  $\Omega$ -stability conjectures for flows”, *Ann. of Math. (2)* **145**:1 (1997), 81–137. MR 98b:58096 Zbl 0871.58067
- [Kelley 1955] J. L. Kelley, *General topology*, Van Nostrand, Toronto, ON, 1955. Reprinted in Grad. Texts in Math. **27**, Springer, New York, 1995. MR 16,1136c Zbl 0066.16604
- [Metzger and Morales 2008] R. Metzger and C. A. Morales, “Sectional-hyperbolic systems”, *Ergodic Theory Dynam. Systems* **28**:5 (2008), 1587–1597. MR 2010g:37045 Zbl 1165.37010
- [Morales 2007] C. A. Morales, “Strong stable manifolds for sectional-hyperbolic sets”, *Discrete Contin. Dyn. Syst.* **17**:3 (2007), 553–560. MR 2008a:37036 Zbl 1137.37015
- [Morales and Pacifico 2002] C. A. Morales and M. J. Pacifico, “Lyapunov stability of  $\omega$ -limit sets”, *Discrete Contin. Dyn. Syst.* **8**:3 (2002), 671–674. MR 2003b:37024 Zbl 1162.37302
- [Morales and Pacifico 2003] C. A. Morales and M. J. Pacifico, “A dichotomy for three-dimensional vector fields”, *Ergodic Theory Dynam. Systems* **23**:5 (2003), 1575–1600. MR 2005a:37030 Zbl 1040.37014
- [Morales et al. 1999] C. A. Morales, M. J. Pacifico, and E. R. Pujals, “Singular hyperbolic systems”, *Proc. Amer. Math. Soc.* **127**:11 (1999), 3393–3401. MR 2000c:37034 Zbl 0924.58068
- [Obata 2010] D. J. Obata, “Resultados na teoria de dinâmica genérica”, undergraduate monograph, Federal University of Rio de Janeiro, 2010, available at <http://im.ufrj.br/~arbiето/davimono.pdf>.
- [Palis 2005] J. Palis, Jr., “A global perspective for non-conservative dynamics”, *Ann. Inst. H. Poincaré Anal. Non Linéaire* **22**:4 (2005), 485–507. MR 2006b:37037 Zbl 1143.37016
- [Palis and de Melo 1982] J. Palis, Jr. and W. de Melo, *Geometric theory of dynamical systems: an introduction*, Springer, New York, 1982. MR 84a:58004 Zbl 0491.58001
- [Shub 1978] M. Shub, *Stabilité globale des systèmes dynamiques*, Astérisque **56**, Société Mathématique de France, Paris, 1978. Translated as *Global stability of dynamical systems*, Springer, New York, 1987. MR 80c:58015 Zbl 0396.58014
- [Wen and Xia 2000] L. Wen and Z. Xia, “ $C^1$  connecting lemmas”, *Trans. Amer. Math. Soc.* **352**:11 (2000), 5213–5230. MR 2001b:37024 Zbl 0947.37018

Received: 2011-12-06

Revised: 2013-12-12

Accepted: 2014-01-08

arbiето@im.ufrj.br

*Instituto de Matemática, Universidade Federal do Rio de Janeiro, P. O. Box 68530, 21945-970 Rio de Janeiro, Brazil*

davi.obata@gmail.com

*Instituto de Matemática, Universidade Federal do Rio de Janeiro, P. O. Box 68530, 21945-970 Rio de Janeiro, Brazil*





# Convergence of the maximum zeros of a class of Fibonacci-type polynomials

Rebecca Grider and Kristi Karber

(Communicated by Kenneth S. Berenhaut)

Let  $a$  be a positive integer and let  $k$  be an arbitrary, fixed positive integer. We define a generalized Fibonacci-type polynomial sequence by  $G_{k,0}(x) = -a$ ,  $G_{k,1}(x) = x - a$ , and  $G_{k,n}(x) = x^k G_{k,n-1}(x) + G_{k,n-2}(x)$  for  $n \geq 2$ . Let  $g_{k,n}$  represent the maximum real zero of  $G_{k,n}$ . We prove that the sequence  $\{g_{k,2n}\}$  is decreasing and converges to a real number  $\beta_k$ . Moreover, we prove that the sequence  $\{g_{k,2n+1}\}$  is increasing and converges to  $\beta_k$  as well. We conclude by proving that  $\{\beta_k\}$  is decreasing and converges to  $a$ .

## 1. Introduction

Let  $\alpha$ ,  $\beta$ , and  $k$  be integers, with  $\alpha \neq 0$ . Consider a Fibonacci-type polynomial sequence given by the recurrence relation  $G_{k,0} = -\alpha$ ,  $G_{k,1} = x - \beta$ , and for  $n \geq 2$ ,

$$G_{k,n}(x) = x^k G_{k,n-1}(x) + G_{k,n-2}(x). \quad (1)$$

We should point out that the classical Fibonacci polynomial sequence  $F_n$  is obtained when  $\alpha = -1$ ,  $\beta = 0$ , and  $k = 1$ . Moreover, the Lucas polynomial sequence  $L_n$  is obtained when  $\alpha = -2$ ,  $\beta = 0$ , and  $k = 1$ . Hoggatt and Bicknell [1973] give explicit forms for the zeros of  $F_n$  and  $L_n$ . Even though finding explicit formulas for other Fibonacci-type polynomial sequences has been a challenge, several results about the properties of the zeros of some specific cases are known. For example, G. Moore [1994] and H. Prodinger [1996] studied the asymptotic behavior of the maximal zeros of  $G_{1,n}$  when  $\alpha = \beta = k = 1$ , and Yu, Wang and He [Yu et al. 1996] generalized Moore's result for  $\alpha = \beta = a$ , where  $a$  is any positive integer. F. Mátyás [1998] studied the same problem for  $\alpha = a$ ,  $a \neq 0$  and  $\beta = \pm a$ . More recently, Wang and He [2004] generalized their previous result for any two integers  $\alpha$  and  $\beta$  with  $\alpha \neq 0$ . We also mention the works of P. E. Ricci [1995] and Mátyás [1998] for boundedness results of the zeros of  $G_{1,n}$ . In addition, Molina and Zeleke [2007; 2009] studied the asymptotic behavior of the zeros of  $G_{k,n}$  when  $\alpha = \beta = 1$  and  $k$  is an arbitrary integer.

*MSC2010:* primary 11B39; secondary 11B37, 30C15.

*Keywords:* Fibonacci polynomial, convergence, zeros, roots.

Moore [1994] proved that when  $\alpha = \beta = k = 1$ , the maximum zeros of the odd-indexed polynomials converge to  $\frac{3}{2}$  from below and the maximum roots of the even-indexed polynomials converge to  $\frac{3}{2}$  from above. In that article, a remark was made about the possibilities of investigating asymptotic behaviors of maximum zeros of other Fibonacci-type polynomial sequences. In [Miller and Zeleke 2013], the first author and Zeleke studied the maximum real zeros of the Fibonacci-type polynomial sequence where  $\alpha = \beta = a$ ,  $a$  is a positive integer, and  $k = 2$ . They provided asymptotic results for the maximum real zeros numerically as well as analytically. We extend those results by allowing  $k$  to be an arbitrary, fixed positive integer. The proof techniques expand those used in [Miller and Zeleke 2013] and [Molina and Zeleke 2009].

Before delving into the technical results, we provide a numerical example to motivate our work.

**Example.** Consider the Fibonacci-type polynomial sequence given by the recurrence relation  $G_{k,0} = -2$ ,  $G_{k,1} = x - 2$ , and for  $n \geq 2$ ,

$$G_{k,n}(x) = x^k G_{k,n-1}(x) + G_{k,n-2}(x).$$

In the context of the generalized Fibonacci-type polynomial sequences we study in this paper, this example corresponds to the case when  $a = 2$ . For a fixed positive integer  $k$  and a natural number  $n$ , let  $g_{k,n}$  represent the maximum real root of the polynomial  $G_{k,n}$ . The first six terms in the sequences of the maximum real roots for  $k = 2$ ,  $k = 3$ , and  $k = 4$  are shown in the following three columns, respectively.

$g_{2,1} = 2$	$g_{3,1} = 2$	$g_{4,1} = 2$
$g_{2,2} \doteq 2.359304086$	$g_{3,2} \doteq 2.190327947$	$g_{4,2} \doteq 2.102374082$
$g_{2,3} \doteq 2.350513611$	$g_{3,3} \doteq 2.188965777$	$g_{4,3} \doteq 2.102149889$
$g_{2,4} \doteq 2.350789278$	$g_{3,4} \doteq 2.188978002$	$g_{4,4} \doteq 2.102150474$
$g_{2,5} \doteq 2.350780807$	$g_{3,5} \doteq 2.188977893$	$g_{4,5} \doteq 2.102150473$
$g_{2,6} \doteq 2.350781067$	$g_{3,6} \doteq 2.188977894$	$g_{4,6} \doteq 2.102150473$

For each sequence, the subsequence created by the odd-indexed (i.e.,  $n$  is odd) maximum real roots is increasing. And, the subsequence created by the even-indexed (i.e.,  $n$  is even) maximum real roots is decreasing. In fact, each of the sequences converge to a real number which is dependent on  $k$ . We call this real number  $\beta_k$ . We should mention  $\beta_k$  is also dependent on our choice of  $a$  and for this example,  $a = 2$ . For the sequences above, we have

$$\beta_2 \doteq 2.350781059, \quad \beta_3 \doteq 2.188977894, \quad \beta_4 \doteq 2.102150473.$$

It is also the case that  $\{\beta_k\}$  converges to 2 and it is not a coincidence that this is the value of  $a$ .

### 2. Formulas

At this time, we introduce a few handy formulas that were established in [Molina and Zeleke 2009]. The formulas in the following lemma allow us to write  $G_{k,n}(x)$  in terms of smaller indexed functions.

**Lemma 2.1.** *For  $n \geq 1$ , the following recursive formulas are true:*

$$G_{k,2n+2}(x) = (x^{2k} + 1)G_{k,2n}(x) + x^{2k}G_{k,2n-2}(x) + \dots + x^{2k}G_{k,2}(x) + x^kG_{k,1}(x),$$

$$G_{k,2n+1}(x) = (x^{2k} + 1)G_{k,2n-1}(x) + x^{2k}G_{k,2n-3}(x) + \dots + x^{2k}G_{k,1}(x) + x^kG_{k,0}(x).$$

The formula that we present in the next lemma provides a type of shift from one indexed polynomial evaluated at  $g_{k,n}$  to another indexed polynomial evaluated at  $g_{k,n}$ . The proof can be found in [Molina and Zeleke 2009, Lemma 4].

**Lemma 2.2.** *For  $n \geq m$ ,  $G_{k,n+m}(g_{k,n}) = (-1)^{m+1}G_{k,n-m}(g_{k,n})$ .*

### 3. Preliminary results

We're now ready to study the maximum real roots,  $g_{k,n}$ , for the generalized Fibonacci-type polynomial sequence defined by  $G_{k,0}(x) = -a$ ,  $G_{k,1}(x) = x - a$ , and  $G_{k,n}(x) = x^kG_{k,n-1}(x) + G_{k,n-2}(x)$  for  $n \geq 2$ , where  $a$  is a positive integer and  $k$  is an arbitrary, fixed positive integer.

**Proposition 3.1.** *If  $n \geq 2$ , then  $g_{k,n} \in (a, a + 1)$ .*

*Proof.* For  $n \geq 2$ , we will show  $G_{k,n}(a) < 0$  and  $G_{k,n}(x) > 0$  for  $x \in [a + 1, \infty)$ ; thus, our conclusion will follow. We'll begin by showing  $G_{k,n}(a) < 0$  by induction. Since  $G_{k,0}(a) = -a$  and  $G_{k,1}(a) = a - a = 0$ , we have  $G_{k,2}(a) = a^k(0) - a = -a < 0$ . Now suppose  $G_{k,m}(a) < 0$  for all  $m$  such that  $2 \leq m \leq n$ . By (1) and the inductive hypothesis,  $G_{k,n+1}(a) = a^kG_{k,n}(a) + G_{k,n-1}(a) < 0$ . Hence,  $G_{k,n}(a) < 0$  for  $n \geq 2$ .

For the remainder of the proof, let  $x \in [a + 1, \infty)$ . We again use induction. Notice

$$G_{k,1}(x) = x - a \geq a + 1 - a > 0, \quad \text{and}$$

$$G_{k,2}(x) = x^k(x - a) - a \geq (a + 1)^k(a + 1 - a) - a = (a + 1)^k - a > 0.$$

Now suppose  $G_{k,m}(x) > 0$  for all  $m$  such that  $2 \leq m \leq n$ . By (1) and the inductive hypothesis, it follows that  $G_{k,n+1}(x) = x^kG_{k,n}(x) + G_{k,n-1}(x) > 0$ . Hence,  $G_{k,n}(x) > 0$  for  $x \in [a + 1, \infty)$  and  $n \geq 2$ .

Therefore,  $g_{k,n} \in (a, a + 1)$  for  $n \geq 2$ . □

**Proposition 3.2.** *Let  $a$  be a positive integer and let  $\beta_k$  be a positive real number that satisfies the equation  $G_{k,2}(x) = -(a - x)^2/a$ ; that is,  $\beta_k$  is a zero of  $T_k(x) = ax^k - a^2x^{k-1} + x - 2a$ . Then*

$$G_{k,n}(\beta_k) = \frac{-(a - \beta_k)^n}{a^{n-1}} \quad \text{for all } n \geq 0.$$

*Proof.* We prove this proposition by induction. The result is true for  $n = 0$  and  $n = 1$  by simple computation. It is true for  $n = 2$  by construction. Now assume  $G_{k,n}(\beta_k) = -(a - \beta_k)^n / a^{n-1}$  for all positive integers less than or equal to  $n$ . Then

$$\begin{aligned}
 G_{k,n+1}(\beta_k) &= \beta_k^k G_{k,n}(\beta_k) + G_{k,n-1}(\beta_k) \\
 &= \beta_k^k \left( \frac{-(a - \beta_k)^n}{a^{n-1}} \right) + \frac{-(a - \beta_k)^{n-1}}{a^{n-2}} \\
 &= \frac{-(a - \beta_k)^{n-1}}{a^{n-2}} \left( \frac{\beta_k^k (a - \beta_k)}{a} + 1 \right) \\
 &= \frac{-(a - \beta_k)^{n-1}}{a^{n-2}} \left( \frac{a\beta_k^k (a - \beta_k) + a^2}{a^2} \right) \\
 &= \frac{-(a - \beta_k)^{n-1}}{a^n} (a\beta_k^k (a - \beta_k) + a^2) \\
 &= \frac{-(a - \beta_k)^{n-1}}{a^n} (-a(\beta_k^k (\beta_k - a) - a)) \\
 &= \frac{-(a - \beta_k)^{n-1}}{a^n} \left( -a \left( \frac{-(a - \beta_k)^2}{a} \right) \right) \\
 &= \frac{-(a - \beta_k)^{n-1}}{a^n} (a - \beta_k)^2 \\
 &= \frac{-(a - \beta_k)^{n+1}}{a^n}.
 \end{aligned}$$

Therefore, our result is true for all nonnegative integers.  $\square$

We remind the reader that whenever  $\beta_k$  is used in this article, it will be dependent on the choice of  $a$ .

**Corollary 3.3.**  $\lim_{n \rightarrow \infty} G_{k,n}(\beta_k) = 0.$

*Proof.* Before we begin, we kindly remind the reader that  $k \geq 1$  and this assumption is continued throughout our work unless stated otherwise. Now the first fact we establish for this proof is that  $\beta_k \in (a, a + 1)$ . To show this, we will again consider  $T_k(x) = ax^k - a^2x^{k-1} + x - 2a$ . It is easily verified that  $T_k(a) < 0 < T_k(a + 1)$ . Moreover,  $T_k$  is strictly increasing on the interval  $[a, \infty)$ , which will be shown by examining the first derivative of  $T_k$ . Notice

$$\begin{aligned}
 T'_k(x) &= kax^{k-1} - (k-1)a^2x^{k-2} + 1 \\
 &= ax^{k-2}(kx - ka + a) + 1 \\
 &= ax^{k-2}(k(x - a) + a) + 1 \\
 &> 0
 \end{aligned}$$

for all  $x \in [a, \infty)$ . Thus,  $\beta_k \in (a, a + 1)$ . Therefore,

$$\lim_{n \rightarrow \infty} G_{k,n}(\beta_k) = \lim_{n \rightarrow \infty} \frac{-(a - \beta_k)^n}{a^{n-1}} = 0. \quad \square$$

#### 4. Analysis of $G'_{k,3}(x)$

In order to prove our main result on the convergence of the maximum zeros, we will need a lower bound on the values  $G'_{k,n}(g_{k,n})$ . This section will provide a lower bound of  $G'_{k,3}(x)$  on the interval  $[g_{k,3}, \infty)$ . We begin with a couple of lemmas to help us achieve this lower bound.

**Lemma 4.1.** *For  $k \geq 3$ ,  $G''_{k,3}(x)$  has exactly one zero in the interval  $(0, \infty)$ .*

*Proof.* Let  $k \geq 3$  and recall  $G_{k,3}(x) = x^{2k+1} - ax^{2k} - ax^k + x - a$ . Thus,

$$\begin{aligned} G''_{k,3}(x) &= (2k + 1)(2k)x^{2k-1} - 2ka(2k - 1)x^{2k-2} - k(k - 1)ax^{k-2} \\ &= kx^{k-2}(2(2k + 1)x^{k+1} - 2a(2k - 1)x^k - a(k - 1)) \\ &= kx^{k-2}f(x), \end{aligned}$$

where  $f(x) = 2(2k + 1)x^{k+1} - 2a(2k - 1)x^k - a(k - 1)$ . We can see that 0 is a zero of  $G''_{k,3}$ . In order to show  $G''_{k,3}$  has only one zero in  $(0, \infty)$ , we will show that  $f(x)$  has exactly one zero in  $(0, \infty)$ . To do so, consider

$$\begin{aligned} f'(x) &= 2(2k + 1)(k + 1)x^k - 2a(2k - 1)kx^{k-1} \\ &= 2x^{k-1}((2k + 1)(k + 1)x - a(2k - 1)k). \end{aligned}$$

The critical numbers of  $f$  are

$$c_1 = 0 \quad \text{and} \quad c_2 = \frac{a(2k - 1)k}{(2k + 1)(k + 1)}.$$

Using this information, it can be verified that  $f$  is decreasing on  $(0, c_2)$  and increasing on  $(c_2, \infty)$ . Pairing this with  $f(0) = -a(k - 1) < 0$  and  $\lim_{x \rightarrow \infty} f(x) = \infty$ , we conclude  $f$ , and hence  $G''_{k,3}$ , has exactly one zero in  $(0, \infty)$ . Therefore, our conclusion holds. □

**Lemma 4.2.** *For  $k \geq 3$ ,  $G'_{k,3}(x)$  has exactly two zeros in the interval  $(0, \infty)$ .*

*Proof.* Let  $k \geq 3$  and recall  $G_{k,3}(x) = x^{2k+1} - ax^{2k} - ax^k + x - a$ . Thus,

$$G'_{k,3}(x) = (2k + 1)x^{2k} - 2kax^{2k-1} - kax^{k-1} + 1.$$

Using the intermediate value theorem and the inequalities  $G'_{k,3}(0) = 1 > 0$ ,  $G'_{k,3}(1) = k(2 - 3a) + 2 \leq -1 < 0$ , and  $\lim_{x \rightarrow \infty} G'_{k,3}(x) = \infty$ , we can conclude  $G'_{k,3}(x)$  has at least two zeros in  $(0, \infty)$ . To show there can be no more than two zeros in  $(0, \infty)$ , we will explore the possibility of  $G'_{k,3}(x)$  having at least three zeros in  $(0, \infty)$ . If

$G'_{k,3}(x)$  has at least three zeros in  $(0, \infty)$ , then  $G''_{k,3}$  would have at least two zeros in  $(0, \infty)$  by Rolle's theorem, but, by Lemma 4.1, we know this cannot be the case. Thus,  $G'_{k,3}(x)$  has exactly two zeros in  $(0, \infty)$  and since  $G'_{k,3}(0) \neq 0$ , those two zeros are indeed in  $(0, \infty)$ .  $\square$

We are now ready to obtain a lower bound on  $G'_{k,3}(x)$  for  $x \in [g_{k,3}, \infty)$ .

**Proposition 4.3.** *If  $k \geq 1$  and  $x \in [g_{k,3}, \infty)$ , then  $G'_{k,3}(x) > 1$ .*

*Proof.* Let  $x \in [g_{k,3}, \infty)$ . We break our proof into cases.

**Case 1:** Consider  $k = 1$ . We then have

- $G_{1,3}(x) = x^3 - ax^2 - ax + x - a$ ,
- $G'_{1,3}(x) = 3x^2 - 2ax - a + 1$ , and
- $G''_{1,3}(x) = 6x - 2a$ .

Since  $G''_{1,3}(x) > 0$  for  $x \in (a/3, \infty)$ , we know  $G'_{1,3}$  is increasing on  $(a/3, \infty)$ . Thus,  $1 \leq G'_{1,3}(a) < G'_{1,3}(x)$  when  $x \in [g_{1,3}, \infty)$  as  $g_{1,3} > a$  by Proposition 3.1.

**Case 2:** Consider  $k = 2$ . We then have

- $G_{2,3}(x) = x^5 - ax^4 - ax^2 + x - a$ ,
- $G'_{2,3}(x) = 5x^4 - 4ax^3 - 2ax + 1$ , and
- $G''_{2,3}(x) = 2(10x^3 - 6ax^2 - a)$ .

Since  $G''_{2,3}(x) > 0$  for  $x \in (a, \infty)$ , we know  $G'_{2,3}$  is increasing on  $(a, \infty)$ . Again notice  $g_{2,3} > a$  by Proposition 3.1. Applying the mean value theorem, we know there exists  $c \in (a, g_{2,3})$  such that

$$G'_{2,3}(c) = \frac{G_{2,3}(g_{2,3}) - G_{2,3}(a)}{g_{2,3} - a}.$$

It follows that when  $x \in [g_{2,3}, \infty)$ ,

$$G'_{2,3}(x) > G'_{2,3}(c) = \frac{G_{2,3}(g_{2,3}) - G_{2,3}(a)}{g_{2,3} - a} = \frac{0 - G_{2,3}(a)}{g_{2,3} - a} = \frac{a^3}{g_{2,3} - a} > 1.$$

**Case 3:** Consider  $k \geq 3$ . By Lemma 4.1, we know  $G''_{k,3}(x)$  has one positive root, call it  $r$ , and, by Lemma 4.2, we know  $G'_{k,3}(x)$  has two positive roots, call them  $s$  and  $t$ , where  $s < t$ . Moreover, by Rolle's theorem,  $s < r < t$ . Notice that

- $G'_{k,3}(0) = 1 > 0$ ,
- $G'_{k,3}(1) = k(2 - 3a) + 2 \leq -1 < 0$ ,
- $\lim_{x \rightarrow \infty} G'_{k,3}(x) = \infty$ , and
- $G''_{k,3}$  is positive on  $(r, \infty)$ .

Thus,  $s < 1 < t$ . Moreover,  $G'_{k,3}$  is negative on  $(s, t)$  and  $G'_{k,3}$  is positive and increasing on  $(t, \infty)$ , and, by the mean value theorem, there exists  $c \in [1, g_{k,3}]$  such that

$$G'_{k,3}(c) = \frac{G_{k,3}(g_{k,3}) - G_{k,3}(1)}{g_{k,3} - 1} = \frac{0 - (2 - 3a)}{g_{k,3} - 1} = \frac{3a - 2}{g_{k,3} - 1} \geq 1.$$

Hence,  $c > t$ , and thus  $g_{k,3} > t$ . Therefore, if  $x \in [g_{k,3}, \infty)$ , then

$$G'_{k,3}(x) > G'_{k,3}(c) \geq 1.$$

Therefore, our conclusion holds for all cases. □

We're now ready to prove that all of the first derivatives of the polynomials are bounded below by 1 as well as explore the characteristics of the maximum zeros. We break this up into two sections, one with the odd-indexed polynomials and the other with the even-indexed polynomials.

### 5. Odd-indexed polynomials

We will use the following two propositions to help establish our results. The proofs are left to the reader as they are similar to those found in [Molina and Zeleke 2009, Lemmas 6 and 7].

**Proposition 5.1.** *The maximum zeros of the odd-indexed polynomials  $G_{k,2n+1}$  form a strictly increasing sequence.*

**Proposition 5.2.** *If  $n \geq 0$ , then the derivative of  $G_{k,2n+1}(x)$  is bounded below by 1 for  $x \in [g_{k,2n+1}, \infty)$ .*

**Proposition 5.3.** *If  $n \geq 0$ , then  $g_{k,2n+1} < \beta_k$  for each  $k \geq 1$ .*

*Proof.* By Proposition 3.2 and for  $n \geq 1$ ,

$$G_{k,2n+1}(\beta_k) = \frac{-(a - \beta_k)^{2n+1}}{a^{2n}} > 0$$

as  $\beta_k \in (a, a + 1)$ . Our goal is to show that

$$G'_{k,2n+1}(x) > G'_{k,2n-1}(x) > \dots > G'_{k,3}(x) > G'_{k,1}(x) = 1$$

for  $x \in [\beta_k, \infty)$  as it will then follow that  $g_{k,2n+1} < \beta_k$ . Now, since  $G_{k,3}(x) \leq 0$  on  $[a, g_{k,3}]$ , it must be the case that  $\beta_k > g_{k,3}$ . Proposition 5.2 gives

$$G'_{k,3}(x) > G'_{k,1}(x) = 1$$

on  $[g_{k,3}, \infty)$ . Thus,

$$G'_{k,3}(x) > G'_{k,1}(x) = 1$$

on  $[\beta_k, \infty)$  as  $[\beta_k, \infty) \subseteq [g_{k,3}, \infty)$ . We note that the rest of the proof follows a similar format to the induction argument used in Proposition 5.2 with  $[\beta_k, \infty)$  replacing  $[g_{k,2n+1}, \infty)$ .  $\square$

## 6. Even-indexed polynomials

**Proposition 6.1.** *If  $n \geq 1$ , then the derivative of  $G_{k,2n}(x)$  is bounded below by 1 for  $x \in [g_{k,2n-1}, \infty)$ .*

*Proof.* We will make use of induction to obtain our result. Let  $x \in [g_{k,2n-1}, \infty)$ . For  $n = 1$ , we have

$$G'_{k,2}(x) = (k+1)x^k - akx^{k-1} = x^{k-1}((k+1)x - ak) > 1.$$

By (1), we have

$$\begin{aligned} G_{k,2n}(x) &= x^k G_{k,2n-1}(x) + G_{k,2n-2}(x), \quad \text{and} \\ G'_{k,2n}(x) &= x^k G'_{k,2n-1}(x) + kx^{k-1} G_{k,2n-1}(x) + G'_{k,2n-2}(x). \end{aligned}$$

From Proposition 5.1, we know  $kx^{k-1} G_{k,2n-1}(x) \geq 0$  as  $x \in [g_{k,2n-1}, \infty)$ . So,

$$G'_{k,2n}(x) \geq x^k G'_{k,2n-1}(x) + G'_{k,2n-2}(x).$$

Now suppose  $G'_{k,2n-2}(x) \geq 1$ . Then

$$\begin{aligned} G'_{k,2n}(x) &\geq x^k G'_{k,2n-1}(x) + G'_{k,2n-2}(x) \\ &> G'_{k,2n-2}(x) \quad (\text{as } x^k G'_{k,2n-1}(x) > 1 \text{ by Proposition 5.2}) \\ &\geq 1 \quad (\text{by the induction hypothesis}). \end{aligned}$$

Therefore, the derivative of the even-indexed polynomials are bounded below by 1 for  $x \in [g_{k,2n-1}, \infty)$ .  $\square$

Referring back to Proposition 5.3, we should note that the result in Proposition 6.1 also holds for  $x \in [\beta_k, \infty)$  as  $[\beta_k, \infty) \subseteq [g_{k,2n-1}, \infty)$ .

**Proposition 6.2.** *The maximum zeros of the even-indexed polynomials form a decreasing sequence that is bounded below by  $\beta_k$ .*

*Proof.* Let  $n \geq 1$ . By Proposition 3.2,

$$G_{k,2n}(\beta_k) = \frac{-(a - \beta_k)^{2n}}{a^{2n-1}} < 0.$$

Thus,  $\beta_k < g_{k,2n}$ . We proceed by induction to show the maximum zeros of the even-indexed polynomials form a decreasing sequence. Notice that

$$G_{k,4}(x) = x^k G_{k,3}(x) + G_{k,2}(x)$$



implies

$$G_{k,4}(g_{k,2}) = g_{k,2}^k G_{k,3}(g_{k,2}) + G_{k,2}(g_{k,2}) = g_{k,2}^k G_{k,3}(g_{k,2}) > 0$$

by utilizing Proposition 5.3. Since  $G_{k,4}$  is increasing on  $[\beta_k, \infty)$  as well, we conclude that  $g_{k,2} > g_{k,4}$ . Now assume  $g_{k,2} > g_{k,4} > \dots > g_{k,2n}$ . By Lemma 2.2,  $G_{k,2n-2}(g_{k,2n}) = -G_{k,2n+2}(g_{k,2n})$ . Since  $g_{k,2n-2} > g_{k,2n}$  (induction hypothesis),  $G_{k,2n-2}$  is increasing on  $[\beta_k, \infty)$ , and  $G_{k,2n-2}(g_{k,2n-2}) = 0$ , it follows that

$$G_{k,2n-2}(g_{k,2n}) < 0 \quad \text{and} \quad G_{k,2n+2}(g_{k,2n}) > 0,$$

and, since  $G_{k,2n+2}(x)$  is increasing on  $[\beta_k, \infty)$ , we have  $g_{k,2n} > g_{k,2n+2}$ . Therefore,  $g_{k,2} > g_{k,4} > \dots > \beta_k$ . □

### 7. Main results

**Theorem 7.1.** *The sequence of odd-indexed zeros is increasing and converges to  $\beta_k$ , and the sequence of even-indexed zeros is decreasing and converges to  $\beta_k$  as well.*

*Proof.* By Proposition 5.1 and Proposition 5.3, we have shown the maximum zeros of the odd-indexed polynomials form an increasing sequence bounded above by  $\beta_k$ , and, by Proposition 6.2, we know the maximum zeros of the even-indexed polynomials form a decreasing sequence bounded below by  $\beta_k$ . In order to show both of the sequences converge to  $\beta_k$ , we will show that  $\lim_{n \rightarrow \infty} g_{k,n} = \beta_k$ . The mean value theorem tells us there exists a real number  $c$  between  $g_{k,n}$  and  $\beta_k$  such that

$$|G'_{k,n}(c)| = \left| \frac{G_{k,n}(\beta_k) - G_{k,n}(g_{k,n})}{\beta_k - g_{k,n}} \right| = \left| \frac{G_{k,n}(\beta_k)}{\beta_k - g_{k,n}} \right|.$$

Since  $G'_{k,n}(c) \geq 1$ ,  $|\beta_k - g_{k,n}| \leq |G_{k,n}(\beta_k)|$ . By utilizing Corollary 3.3, which states  $\lim_{n \rightarrow \infty} G_{k,n}(\beta_k) = 0$ , we can say  $\lim_{n \rightarrow \infty} g_{k,n} = \beta_k$ . Therefore, the sequence of odd-indexed zeros and the sequence of even-indexed zeros converge to  $\beta_k$ . □

**Theorem 7.2.** *The sequence  $\{\beta_k\}$  is decreasing and converges to  $a$ .*

*Proof.* We begin by referring the reader back to  $T_k(x)$  as defined in Proposition 3.2. Recall that  $T_k$  is increasing on  $[a, \infty)$  and  $\beta_k \in (a, a + 1)$  is a zero of  $T_k$ . Using the fact that  $\beta_k$  is a zero of  $T_k$ , we have  $a\beta_k^k - a^2\beta_k^{k-1} = 2a - \beta_k$ . Then

$$\begin{aligned} T_{k+1}(\beta_k) &= a\beta_k^{k+1} - a^2\beta_k^k + \beta_k - 2a = \beta_k(a\beta_k^k - a^2\beta_k^{k-1}) + \beta_k - 2a \\ &= \beta_k(2a - \beta_k) + \beta_k - 2a = (\beta_k - 1)(2a - \beta_k) \\ &> 0. \end{aligned}$$

Thus,  $\beta_{k+1} < \beta_k$ , which verifies that  $\{\beta_k\}$  is decreasing. Now let  $\varepsilon > 0$ . Then

$$\begin{aligned}
\lim_{k \rightarrow \infty} T_k(a + \varepsilon) &= \lim_{k \rightarrow \infty} [a(a + \varepsilon)^k - a^2(a + \varepsilon)^{k-1} + (a + \varepsilon) - 2a] \\
&= \lim_{k \rightarrow \infty} [a(a + \varepsilon)^{k-1}(a + \varepsilon - a) + a + \varepsilon - 2a] \\
&= \lim_{k \rightarrow \infty} [\varepsilon a(a + \varepsilon)^{k-1} + \varepsilon - a] \\
&= \infty.
\end{aligned}$$

We then know that there exists  $j \in \mathbb{Z}$  such that  $T_j(a + \varepsilon) > 0$  and so  $\beta_j \in (a, a + \varepsilon)$ . Therefore,  $\lim_{k \rightarrow \infty} \beta_k = a$ .  $\square$

### Acknowledgements

The authors would like to thank and acknowledge A. Zeleke for his introduction to this research topic via an REU program at Michigan State University.

### References

- [Hoggatt and Bicknell 1973] V. E. Hoggatt, Jr. and M. Bicknell, “Roots of Fibonacci polynomials”, *Fibonacci Quart.* **11**:3 (1973), 271–274. MR 48 #2056
- [Mátyás 1998] F. Mátyás, “Bounds for the zeros of Fibonacci-like polynomials”, *Acta Acad. Paedagog. Agriensis Sect. Mat. (N.S.)* **25** (1998), 15–20. MR 2000h:11015
- [Miller and Zeleke 2013] R. Miller (R. Grider) and A. Zeleke, “On the zeros of Fibonacci type polynomials with varying initial conditions”, *Congr. Numer.* **216** (2013), 109–117.
- [Molina and Zeleke 2007] R. Molina and A. Zeleke, “On the convergence of the maximum roots of a Fibonacci-type polynomial sequence”, *Congr. Numer.* **184** (2007), 121–128. MR 2009a:11032
- [Molina and Zeleke 2009] R. Molina and A. Zeleke, “Generalizing results on the convergence of the maximum roots of Fibonacci type polynomials”, *Congr. Numer.* **195** (2009), 95–104. MR 2584288
- [Moore 1994] G. A. Moore, “The limit of the golden numbers is  $3/2$ ”, *Fibonacci Quart.* **32**:3 (1994), 211–217. MR 95f:11008
- [Prodinger 1996] H. Prodinger, “The asymptotic behavior of the golden numbers”, *Fibonacci Quart.* **34**:3 (1996), 224–225. MR 97d:11032
- [Ricci 1995] P. E. Ricci, “Generalized Lucas polynomials and Fibonacci polynomials”, *Riv. Mat. Univ. Parma* (5) **4** (1995), 137–146. MR 97b:11018
- [Wang and He 2004] Y. Wang and M. He, “Zeros of a class of Fibonacci-type polynomials”, *Fibonacci Quart.* **42**:4 (2004), 341–347. MR 2005h:11035
- [Yu et al. 1996] H. Yu, Y. Wang, and M. He, “On the limit of generalized golden numbers”, *Fibonacci Quart.* **34**:4 (1996), 320–322. MR 97b:11020

Received: 2012-10-07    Revised: 2013-06-16    Accepted: 2013-10-19

Rebecca.Miller-1@ou.edu

*Department of Mathematics, University of Oklahoma, 601 Elm Avenue, Room 423, Norman, OK 73019, United States*

kkarber1@uco.edu

*Department of Mathematics and Statistics, University of Central Oklahoma, 100 North University Drive, Edmond, OK 73034, United States*

# Iteration digraphs of a linear function

Hannah Roberts

(Communicated by Robert W. Robinson)

An iteration digraph  $G(n)$  generated by the function  $f(x) \bmod n$  is a digraph on the set of vertices  $V = \{0, 1, \dots, n - 1\}$  with the directed edge set  $E = \{(v, f(v)) \mid v \in V\}$ . Focusing specifically on the function  $f(x) = 10x \bmod n$ , we consider the structure of these graphs as it relates to the factors of  $n$ . The cycle lengths and number of cycles are determined for various sets of integers including powers of 2 and multiples of 3.

## 1. Introduction

Using the graph  $D_7$ , shown in Figure 1, the remainder modulo 7 of any integer  $N$  can be determined based solely on the digits of the  $N$  [Wilson 2009]. For example, consider  $N = 375$ . Begin at the vertex labeled 0. First, follow three black edges. Then follow one red edge and seven black edges, ending on 2. Finally, follow one red edge and five black edges to end on 4. This indicates that  $375 \equiv 4 \pmod{7}$ .

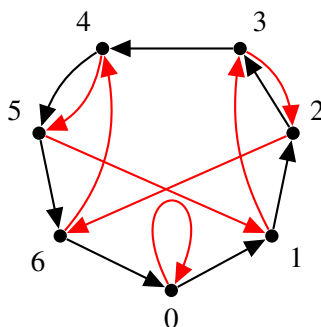
Generalizing this algorithm to any  $N$  where  $d_i$  is the  $i$ -th digit, we start at 0 and follow  $d_1$  black edges. We then continue to follow  $d_i$  black edges for  $i = 2, 3, \dots, r$ . Between each digit, we follow one red edge. The vertex where we end after the final  $d_r$  black edges is the remainder when  $N$  is divided by 7.

The graph  $D_7$  is formed by two specific iteration digraphs, directed graphs each generated by a function  $f : \mathbb{Z}_n \rightarrow \mathbb{Z}_n$ . The graph  $G_n$  is formed on the vertex set  $V = \mathbb{Z}_n = \{0, 1, 2, \dots, n - 1\}$  with exactly one edge from  $v$  to  $f(v)$  for all  $v \in V$ . Thus, the edge set is  $E = \{(v, f(v)) \mid v \in V\}$ , where  $(v, f(v))$  indicates the edge directed from vertex  $v$  to  $f(v)$ . The red edges in  $D_7$  form the iteration digraph produced by the function  $f(x) \equiv 10x \pmod{7}$ . Thus,  $V(D_7) = \{0, 1, 2, \dots, 6\}$ , and  $E(D_7)$  includes  $(1, 3)$ ,  $(3, 2)$ , and so on, because  $10 \equiv 3 \pmod{7}$  and  $30 \equiv 2 \pmod{7}$ . The black edges are generated by the function  $g(x) \equiv x + 1 \pmod{7}$ .

Using these two functions, divisibility graphs can easily be drawn for any integer  $n$ , and the same algorithm will produce remainders modulo  $n$ . Given this, one may naturally question how the graph produced by  $f(x) \bmod n$  changes for

MSC2010: 05C20, 11A07.

Keywords: digraph, cycle, congruence.



**Figure 1.** The graph  $D_7$ , used to determine divisibility by 7.

different integers  $n$ . This work considers the number and length of the cycles in the graph  $G(n)$  generated by the function  $f(x) = 10x \bmod n$ .

## 2. Relatively prime integers

To begin, we look at the common structures found in a broad subset, the set of all integers relatively prime to 10. The most basic feature of these graphs is given in Theorem 1 below.

A vertex  $v$  in  $G(n)$  is said to be in level  $i$  if the longest path ending at  $v$  which does not contain any part of a cycle has length  $i$  [Sommer and Křížek 2004]. If the highest level vertex in  $G(n)$  is at level  $i$ , then  $G(n)$  has  $i + 1$  levels. Thus,  $G(28)$  (Figure 7) has 3 levels. Level 0 contains 7 and 9, level 1 contains 6, and level 2 contains 0. Also, the indegree of a vertex  $v$ , written  $\text{indeg}(v)$ , is the number of edges directed towards  $v$ . In  $G(28)$ ,  $\text{indeg}(7) = 0$  while  $\text{indeg}(6) = 2$ .

**Theorem 1.**  $G(n)$  has 1 level for all  $n$  with  $\text{gcd}(10, n) = 1$ .

*Proof.* Because  $V(G(n))$  is the complete reduced residue set of  $n$  and  $\text{gcd}(10, n) = 1$ , the set  $S = \{10v \mid v \in V(G(n))\}$  is also a complete residue set [Rosen 2000]. Thus,  $f : V(G(n)) \rightarrow V(G(n))$  is one-to-one and onto, so every vertex has indegree exactly 1.

Now assume  $v \in V(G(n))$  is at level  $i > 0$ . Then there must be a path of  $i$  edges leading to  $v$  which is not part of a cycle. The first vertex in this noncyclic path must have an indegree of 0. This is a contradiction, so  $v$  must be at level 0 and  $G(n)$  has 1 level.  $\square$

The above theorem could be restated to say every vertex in  $G(n)$  is at level 0. From this fact, it is clear that every graph  $G(n)$  with  $\text{gcd}(10, n) = 1$  is simply a set of isolated cycles. That is,  $G(n)$  is a set of cycles without any adjacent noncyclic vertices. We next consider the lengths of these cycles.

The length of the cycles in  $G(n)$  is dependent on the prime factors of  $n$ , but before considering the total number of cycles, we first look at a subset of the vertices.

A graph  $H$  is called a subgraph of  $G$  if  $V(H) \subset V(G)$  and  $E(H) \subset E(G)$ , where the edges in  $E(H)$  must connect vertices in  $V(H)$ . We say  $H$  is generated by  $V(H)$  if  $E(H)$  contains every edge in  $G$  that connects vertices in  $V(H)$ .

**Theorem 2.** *In  $G(n)$ , if  $V_1$  is the subset of vertices relatively prime to  $n$ , then there are  $\phi(n)/\text{ord}_n(10)$  cycles, each of length  $\text{ord}_n(10)$ , in the subgraph generated by  $V_1$ .*

*Proof.* First, let  $(a, b)$  be an edge in  $G(n)$ . Since  $\gcd(10, n) = 1$ , if  $\gcd(a, n) = 1$ , then  $10a \equiv b \pmod{n}$  is also relatively prime to  $n$ . Thus, if a cycle contains one vertex that is relatively prime to  $n$ , then all vertices in the cycle must also be relatively prime to  $n$ .

Now, let  $r = \text{ord}_n(10)$ , so  $r$  is the least integer for which  $10^r \equiv 1 \pmod{n}$ , or equivalently  $10^r v \equiv v \pmod{n}$  for every  $v \in V(G(n))$ . In the sequence of vertices  $\{v_0, v_1, v_2, \dots, v_r\}$  from  $G(n)$ ,  $v_t \equiv 10^t v_0$ . Thus,  $v_r \equiv 10^r v_0 \equiv v_0$  and the sequence is an  $r$ -cycle.

Consider  $s > r$ . We can write  $s = mr + t$ , where  $m, t$ , and  $s$  are integers such that  $0 \leq t < r$ . Since  $10^s v_0 \equiv 10^t v_0 \equiv v_t$ , a path longer than  $r$  will repeat through the cycle. Thus, the longest possible cycle in  $G(n)$  has length  $r$ .

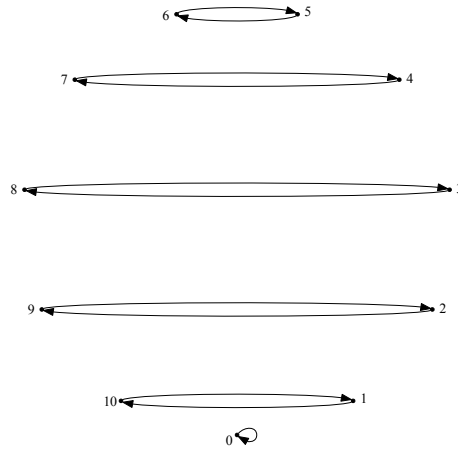
Now, let  $v \in G(n)$  such that  $\gcd(v, n) = 1$ , and assume  $v$  is part of an  $s$ -cycle where  $s < r = \text{ord}_n(10)$ . Then  $10^s v \equiv v \pmod{n}$ , but  $10^s \not\equiv 1 \pmod{n}$ , because by definition  $r$  is the smallest positive integer for which  $10^r \equiv 1 \pmod{n}$ . This means  $10^s - 1 = np + t$  for some integers  $p$  and  $0 < t < n$ . Also,  $10^s v - v = nm$  for some integer  $m$ , so

$$\begin{aligned} v(10^s - 1) &= nm \\ v(np + t) &= nm \\ vt &= n(m - vp). \end{aligned}$$

Now we have  $n \mid (vt)$ , but  $n \nmid t$  because  $0 < t < n$ . Hence,  $\gcd(n, v) > 1$ , which is a contradiction since we assumed  $\gcd(n, v) = 1$ . Therefore, all cycles on vertices relatively prime to  $n$  have length  $r = \text{ord}_n(10)$ . Also, there are  $\phi(n)$  vertices relatively prime to  $n$ , so there are  $\phi(n)/\text{ord}_n(10)$  such cycles.  $\square$

As an example of Theorem 2, consider  $G(11)$  (Figure 2). There are 10 vertices relatively prime to 11,  $V_1 = \{1, 2, 3, \dots, 10\}$ , and  $\text{ord}_{11}(10) = 2$ . Thus,  $G(11)$  contains  $10/2 = 5$  cycles all of length 2.

Define  $C_n$  to be the number of cycles and  $L_n$  to be the set of all cycle lengths in  $G(n)$ . Now the above theorem is used to help determine  $C_n$  and  $L_n$  for any  $n$  relatively prime to 10.



**Figure 2.**  $G_{11}$  contains five 2-cycles.

**Theorem 3.** *Let  $\gcd(10, n) = 1$ . Then*

$$C_n = \sum_{d|n} \frac{\phi(d)}{\text{ord}_d(10)},$$

*and the set of cycle lengths is  $L_n = \{\text{ord}_d(10) \mid d \mid n\}$ .*

*Proof.* First, define the set  $V_d = \{v \in V(G(n)) \mid \gcd(v, n) = d\}$  for all  $d \mid n$ . Every  $v$  in  $G(n)$  will be in exactly one set  $V_d$ , so these sets form a partition of  $V(G(n))$ . Also, define  $G_d(n)$  to be the subgraph of  $G(n)$  generated by the vertex set  $V_d$ .

Let  $a \in V_d$  and  $(a, b) \in E(G(n))$ . Then by reasoning similar to that used in the previous theorem,  $b \in V_d$ .

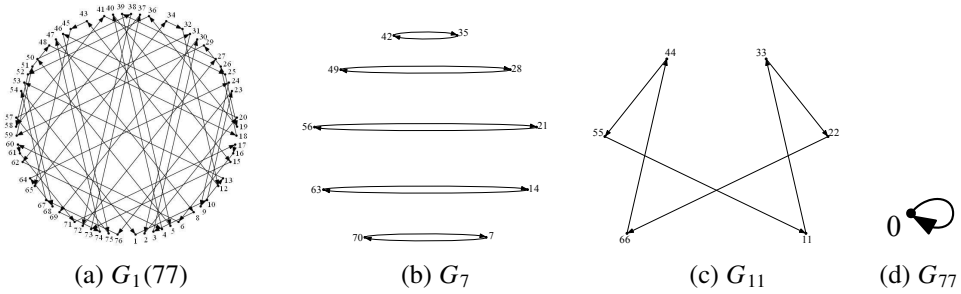
Thus, every cycle in  $G(n)$  contains vertices from exactly one set  $V_d$ , and we can determine  $C_n$  by adding the number of cycles in  $G_d(n)$  for every  $d \mid n$ , or

$$C_n = \sum_{d|n} (\text{number of cycles in } G_d(n)). \tag{1}$$

We now need to find the number of cycles in each subgraph  $G_d(n)$ . Let  $(a, b)$  be an edge in  $G_d(n)$ . We already have  $a = dt$ , where  $\gcd(n/d, t) = 1$ , and similarly,  $b = ds$ , where  $\gcd(n/d, s) = 1$ . Thus,  $(a, b) = (dt, ds)$ . Now,

$$\begin{aligned} 10a - b &= n(p) \\ 10(dt) - ds &= n(p) \\ 10t - s &= \frac{n}{d}(p), \end{aligned}$$

so  $(t, s)$  is an edge in  $G(n/d)$ . Since  $t$  and  $s$  are relatively prime to  $n/d$ , our problem is now equivalent to finding the number of cycles on the vertices of  $G(n/d)$  relatively



**Figure 3.** The subgraphs of  $G(77)$  generated by  $V_1, V_7, V_{11},$  and  $V_{77}$ .

prime to  $n/d$ . In other words, the number of cycles in  $G_d(n)$  is the same as the number of cycles in  $G_1(n/d)$ . From Theorem 2, we know that  $G_1(n/d)$  contains  $\phi(n/d)/\text{ord}_{n/d}(10)$  cycles with length  $\text{ord}_{n/d}(10)$ .

Thus, there are also  $\phi(n/d)/\text{ord}_{n/d}(10)$  cycles in  $G_d(n)$  with length  $\text{ord}_{n/d}(10)$ . Therefore,

$$C_n = \sum_{d|n} \frac{\phi(n/d)}{\text{ord}_{n/d}(10)}.$$

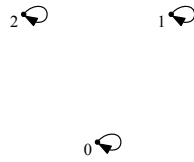
Every divisor  $d_1$  can be written as  $d_1 = n/d_2$  for some other divisor  $d_2$ . Hence, as we sum over every divisor  $d$ , we are also summing over  $n/d$  for every  $d$ , so we can rewrite  $C_n$  as

$$C_n = \sum_{d|n} \frac{\phi(d)}{\text{ord}_d(10)}. \tag{2}$$

This concludes the proof. □

One example of the previous theorem is  $G(77)$  (Figure 3). To make it easier to see the various cycles of  $G(77)$ , Figure 3 shows the subgraphs of  $G(77)$  generated by  $V_d$  for  $d = 1, 7, 11, 77$ . Looking at  $G_{11}(77)$  in Figure 3(c), the vertices all have  $\text{gcd}(v, 77) = 11$ . If we compare this subgraph to  $G(7)$  in Figure 1, we see that  $G_{11}(77)$  is isomorphic to  $G_1(7)$  by the isomorphism  $h(v) = 11v$ . This isomorphism illustrates the relation of edges in  $G(n)$  and in  $G(mn)$ . Similarly,  $G_7(77)$  is isomorphic to  $G_1(11)$ . Finally,  $G_{77}(77)$  in Figure 3(d) is simply the isolated fixed point isomorphic to  $G(1)$  that appears in every  $G(n)$  where  $(10, n) = 1$ .

The isomorphisms seen in  $G(77)$  can be generalized to other  $G(n)$ . For  $d | n$ , the subgraph  $G_d(n)$  is isomorphic to the subgraph  $G_1(n/d)$ . Thus, much of  $G(n)$  is built from the graphs of  $G(d)$ . The subgraph  $G_1(n)$  on the vertices that are relatively prime to  $n$  is the only portion of the total graph  $G(n)$  that can not be built directly from a graph  $G(d)$  for some  $d | n$ .



**Figure 4.** Every vertex in  $G(3)$  is an isolated fixed point.

We now have the basic structure of the graph for any  $n$  relatively prime to 10, and can consider which integers produce a more specific structure. The next section explores how multiples of 3 affect the structure of a graph to produce a set of isomorphic subgraphs.

### 3. Multiples of 3

Because  $10 \equiv 1 \pmod 3$ , for every vertex  $v$  in  $G(3)$ ,  $(v, v)$  is an edge for all  $v \in \{0, 1, 2\}$  (Figure 4). This property of  $G(3)$  leads to a highly predictable structure for  $G(3n)$  when  $\gcd(3, n) = 1$ .

We first need to establish some notation for the vertices of  $G(n)$  and  $G(3n)$ . Define  $V$  to be the vertex set of  $G(n)$ , so  $V = V(G(n)) = \{0, 1, 2, \dots, n - 1\}$ . Also, define

$$V_t = \{3v + tn \pmod{3n} \mid v \in V\} \quad \text{for } t = 0, 1, 2.$$

If  $v \in V$ , then  $v_t = 3v + tn \pmod{3n} \in V_t$ . For  $n = 2$ , we have  $G(2)$  with  $V = \{0, 1\}$  and  $G(3n) = G(6)$  with  $V_0 = \{0, 3\}$ ,  $V_1 = \{2, 5\}$ , and  $V_2 = \{1, 4\}$ , as in Figure 5.

The following theorem uses these vertex sets to relate the edge sets of  $G(n)$  and  $G(3n)$  for  $\gcd(3, n) = 1$ .

**Theorem 4.** *If  $3 \nmid n$  and  $E(G(n)) = \{(a, b) \mid b = f(a), a \in V\}$ , then  $E(G(3n)) = \{(a_t, b_t) \mid (a, b) \in E(G(n)), t = 0, 1, 2\}$ .*

*Proof.* Let  $(a, b)$  be an edge in  $G(n)$ . Thus  $10a \equiv b \pmod n$  and  $3a \equiv 3b \pmod{3n}$ . Considering  $a_t$ ,

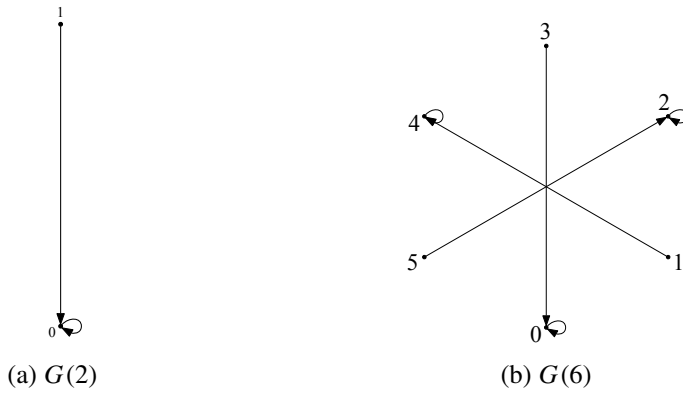
$$\begin{aligned} 10(3a + tn) &\equiv 30a + 10tn \pmod{3n} \\ &\equiv 3b + tn + 3n(3t) \pmod{3n} \\ &\equiv 3b + tn \pmod{3n}. \end{aligned}$$

Therefore,  $(a_t, b_t)$  is also an edge in  $G(3n)$ . We now have that

$$S = \{(a_t, b_t) \mid (a, b) \in E(G(n)), t = 0, 1, 2\}$$

is a subset of  $E(G(3n))$ . By definition of an iteration digraph, we know that  $G(3n)$  has  $3n$  distinct edges. The set  $S$  has  $3n$  edges, which we now need to show are distinct.





**Figure 5.** The components of  $G(6)$  are all isomorphic to  $G(2)$ .

For any  $v, w \in V$ , if  $v \not\equiv w \pmod n$ , then  $v_t \not\equiv w_t \pmod{3n}$ . Hence,  $V_0, V_1$ , and  $V_2$  each contain  $n$  incongruent integers.

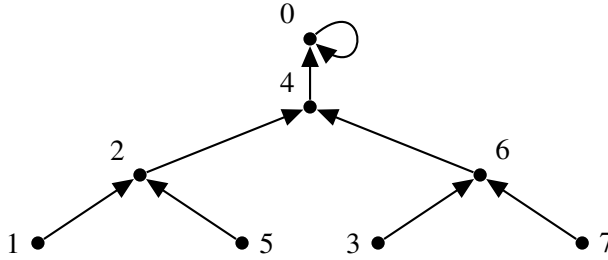
Next, if  $a \in V$ , we have  $a_0 \equiv 0 \pmod 3$ ,  $a_1 \equiv n \pmod 3$ , and  $a_2 \equiv 2n \pmod 3$ . Hence, for any  $b, c, d \in V$ , not necessarily distinct,  $b_0, c_1$ , and  $d_2$  are incongruent modulo 3. Now, assume  $b_r \equiv c_t \pmod{3n}$ , so  $b_r - c_t = 3n(p)$  for some integer  $p$ . Then  $b_r - c_t = 3(np)$  and  $b_r \equiv c_t \pmod 3$ . This is a contradiction since  $b_r$  and  $c_t$  are incongruent mod 3. Hence,  $b_r \not\equiv c_t \pmod{3n}$ . Thus,  $b_0, c_1$ , and  $d_2$  are all incongruent modulo  $3n$ . Furthermore,  $a_t \not\equiv b_r \pmod{3n}$  whenever either  $a \not\equiv b \pmod n$  or  $r \neq t$ . Therefore, the  $3n$  edges in  $S$  are distinct, so  $E(G(3n)) = S = \{(a_t, b_t) \mid (a, b) \in E(G(n)), t = 0, 1, 2\}$ .  $\square$

An example of Theorem 4 is the graphs for  $n = 6$  shown in Figure 5(b). The graph  $G(6)$  has three components on the sets of vertices  $\{0, 3\}$ ,  $\{1, 4\}$ , and  $\{2, 5\}$ . Comparing these to  $G(2)$ , each component is isomorphic to  $G(2)$ . Thus, the relation from Theorem 4 between any  $G(n)$  and  $G(3n)$  can also be expressed in terms of isomorphisms between the graphs.

**Corollary 1.**  $G(3n)$  is the union of three subgraphs, each of which is isomorphic to  $G(n)$ .

A theorem similar to Theorem 4 can be proved for  $G(9n)$  when  $\gcd(3, n) = 1$ . This indicates that perhaps this type of edge relation will exist for higher powers of 3 as well. However, for 3 and 9, the proofs are contingent on the fact that  $10 \equiv 1$  modulo both 3 and 9. Theorem 4 cannot be generalized for  $G(3^k n)$  where  $k \geq 3$ .

Based on Theorem 4, it is also clear that  $G(3n)$  contains exactly 3 times as many cycles as  $G(n)$  with all the same cycle lengths. Thus, while Theorem 3 holds for multiples of 3, we can now say  $C_{3n} = 3C_n$  and  $L_{3n} = L_n$  when  $\gcd(3, n) = 1$ . Similarly,  $C_{9n} = 9C_n$  and  $L_{9n} = L_n$ .



**Figure 6.**  $G(8)$ .

### 4. Powers of 2

Another class of integers for which  $G(n)$  has a distinctive and predictable digraph is the powers of 2. When  $n = 2^k$  for some integer  $k > 0$ ,  $G(2^k)$  takes the form of a binary tree with all edges heading towards the root. This unique form follows from the fact that 2 is a factor of 10. In this section, congruences should all be considered modulo  $2^k$  unless otherwise specified.

Given this tree structure, which will be proved in Theorem 5, each vertex will be referenced by its level and its position within that level. Number the vertices in level  $i < k$  left to right from 0 to  $2^s - 1$ , where  $s = k - i - 1$ . Then  $v_{i,t}$  is the vertex in level  $0 \leq i \leq k$  at position  $0 \leq t \leq 2^s - 1$ . In Figure 6, for example,  $v_{0,0} = 1$ ,  $v_{0,1} = 5$ , and  $v_{1,0} = 2$ . Additionally, for each pair of vertices  $v_{i,t}$  and  $v_{i,t+1}$  where both are adjacent to the same vertex at level  $i + 1$ , we will draw the graph such that  $v_{i,t} < v_{i,t+1}$ .

We can now develop the basic structure of the  $2^k$  iteration digraph.

**Theorem 5.** *If  $G(n)$  is the iteration digraph of  $f(x) \equiv 10x \pmod{2^k}$ , where  $n = 2^k$  for  $k = 1, 2, 3, \dots$ , then:*

- (i)  $G(n)$  has  $k + 1$  levels.
- (ii) The nonzero vertices form a complete binary tree with height  $k$ .
- (iii) Exactly 2 vertices at level  $i < k - 1$  are adjacent to each vertex at level  $i + 1$ .
- (iv) For each vertex  $v_{i,t}$  at level  $i < k$ ,  $2^i \parallel v_{i,t}$ .

*Proof.* For part (i), we know for any vertex  $v$  that  $10^k v = 2^k (5^k v) \equiv 0 \pmod{2^k}$ . Thus, the longest possible path from  $v$  to 0 has length  $k$ . Now suppose the longest path that exists is only  $k - 1$  edges long. Then  $10^{k-1} v = 2^{k-1} (5^{k-1} v) \equiv 0$  for all  $v$ . This means that

$$2^{k-1} (5^{k-1} v) = 2^k p$$

$$5^{k-1} v = 2p,$$

and  $v$  must be divisible by 2. This is a contradiction for all odd vertices, so there must exist a path from  $v$  to 0 with length  $k$ . Thus,  $G(2^k)$  has  $k + 1$  levels.

Considering part (iv), at level  $k - 1$ , we have  $2^{k-1} \parallel 2^{k-1}$ . Now, for induction down the levels, assume that  $2^i \parallel v_{i,t}$  for all vertices at some level  $i \leq k - 1$  and let  $v_{i-1,r}$  be adjacent to  $v_{i,t} = 2^i c$ , where  $c$  is an odd integer. Hence,  $v_{i-1,r}$  is at level  $i - 1$  and

$$\begin{aligned} 10v_{i-1,r} - v_{i,t} &= 2^k b \\ 10v_{i-1,r} &= 2^i (2^{k-i} b + c). \end{aligned}$$

Thus,  $2^i$  divides  $10v_{i-1,r}$ , so  $2^{i-1}$  divides  $v_{i-1,r}$ .

We now need to show that  $2^{i-1} \parallel v_{i-1,r}$ . Assume that  $2^i \mid v_{i-1,r}$ . Then  $10v_{i-1,r} \equiv v_{i,t}$  is divisible by  $2^{i+1}$ . This is a contradiction to the initial assumption that  $2^i \parallel v_{i,t}$ . Therefore,  $2^i$  does not divide  $v_{i-1,r}$ , so  $2^{i-1} \parallel v_{i-1,r}$ , and for every vertex  $v_{i,t}$  at a level  $i < k$ ,  $2^i \parallel v_{i,t}$ .

For part (iii), let  $a$  and  $b$  be vertices such that  $f(a) = b$  and  $b$  is at level  $i$ , where  $0 < i \leq k - 1$ . Then consider  $a + 2^{k-1}$ .

$$10(a + 2^{k-1}) \equiv b + 5 \cdot 2^k \equiv b + 0 \pmod{2^k}. \tag{3}$$

Since  $2^{k-1} < 2^k$ ,  $a \not\equiv a + 2^{k-1} \pmod{2^k}$ . Thus, at least two distinct vertices are adjacent to  $b$ . From part (iv), there are  $2^{k-i-1}$  vertices at level  $i$  and  $2^{k-i}$  at level  $i + 1$ , so there are exactly twice as many vertices at level  $i$  as at level  $i + 1$ . Thus, exactly two vertices are adjacent to each vertex at level  $0 < i < k$ .

Part (ii) also follows directly from parts (iii) and (i) and the definition of a tree, so the nonzero vertices form a complete binary tree with height  $k$  and with  $2^{k-1}$  as the root. □

From the above theorem,  $G(2^k)$  can be drawn for any  $k \geq 1$  and we have some idea of the label placement within that graph. It is also clear that  $G(2^k)$  always contains exactly one 1-cycle.

Since  $G(2^k)$  is really just  $G(2^k n)$  with  $n = 1$ , we now consider the more general  $G(2^k n)$  with  $\gcd(10, n) = 1$ . First, we find that  $G(2^k n)$  is semiregular; that is, each vertex in  $G(2^k n)$  has an indegree of either 0 or  $d$ , for some positive integer  $d$ .

**Theorem 6.** *If  $n$  is not divisible by 2 or 5, then  $G(2^k n)$  is semiregular with  $d = 2$  and  $\text{indeg}(v) = 2$  if and only if  $2 \mid v$ .*

*Proof.* Let  $(a, b)$  be an edge in  $G(2^k n)$ . Then  $10a \equiv b \pmod{2^k n}$ , and also

$$\begin{aligned} 10(a + 2^{k-1} n) &\equiv 10a + 5 \cdot 2^k n \pmod{2^k n} \\ 10(a + 2^{k-1} n) &\equiv b + 0 \pmod{2^k n}. \end{aligned} \tag{4}$$

Since  $2^{k-1} n < 2^k n$ ,  $a \not\equiv a + 2^{k-1} n$  and  $(a + 2^{k-1} n, b)$  is also an edge in  $G(2^k n)$ . Thus, if  $\text{indeg}(v) \geq 1$  for any  $v \in V(G(2^k n))$ , then  $\text{indeg}(v) \geq 2$ .

Now, assume there exists a third vertex  $c$  which is also adjacent to  $b$  and is incongruent to both  $a$  and  $a + 2^{k-1}n$ . Then

$$10c - b = 2^k ns \quad \text{and} \quad 10a - b = 2^k np, \tag{5}$$

where  $s$  and  $p$  are integers such that  $s \neq p$ .

From (5) we get

$$\begin{aligned} 10(c - a) &= 2^k n(s - p) \\ 5(c - a) &= 2^{k-1} n(s - p). \end{aligned}$$

Then 5 divides  $(s - p)$ , so  $(s - p) = 5t$  for some nonzero integer  $t$  and

$$\begin{aligned} 5(c - a) &= 2^{k-1} n(5t) \\ c &= a + 2^{k-1} nt. \end{aligned} \tag{6}$$

If  $t$  is even, then  $t = 2r$  and  $c \equiv a + 2^k nr \equiv a \pmod{2^k n}$ . If  $t$  is odd, then  $t = 2r + 1$  and

$$c \equiv a + 2^{k-1} n(2r + 1) \equiv a + 2^{k-1} n \pmod{2^k}.$$

Thus,  $c$  is congruent to either  $a$  or  $a + 2^{k-1}n$ , so the indegree of  $b$  is exactly 2 and the indegree of any vertex of  $G(2^k n)$  is either 0 or 2. Therefore,  $G(2^k n)$  is semiregular with  $d = 2$ .

Now, assume  $(a, b)$  is an edge where  $2 \nmid b$ . Then  $10a \equiv b \pmod{2^k n}$ , so

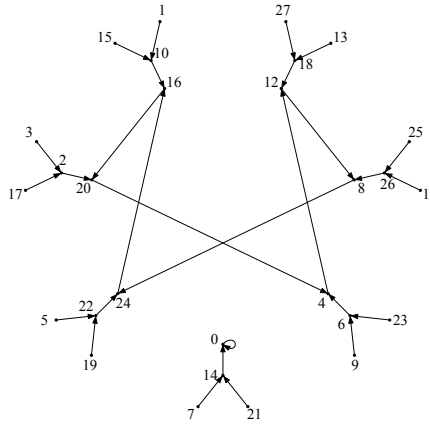
$$\begin{aligned} 10a - b &= 2^k np \\ 10a - 2^k np &= b \\ 2(5a - 2^{k-1} np) &= b. \end{aligned}$$

Thus,  $2 \mid b$ , which is a contradiction, so when  $2 \nmid v$ ,  $\text{indeg}(v) = 0$ . There are  $2^{k-1}n$  vertices that are divisible by 2 and, hence, can have an indegree of 2. Since there are exactly twice as many edges as there are vertices divisible by 2,  $\text{indeg}(v) = 2$  whenever  $2 \mid v$ . Therefore,  $\text{indeg}(v) = 2$  if and only if  $2 \mid v$ .  $\square$

The graph  $G(28)$  is seen to be semiregular with  $d = 2$  in Figure 7. It also includes several subgraphs with a binary tree structure. These subgraphs are isomorphic to  $G(2^2)$ . In the following theorem, these subgraphs isomorphic to  $G(2^k)$  are shown to be present in  $G(2^k n)$  for any  $k \geq 1$  and  $n$  relatively prime to 10.

**Theorem 7.** *If  $n$  is not divisible by 2 or 5 and  $k > 0$ , then  $G(2^k n)$  contains  $n$  generated subgraphs that are isomorphic to the subgraph of  $G(2^k)$  excluding the loop  $(0, 0)$ . The root of each isomorphic subgraph is a vertex  $v \in V(G(2^k n))$ , where  $2^k \mid v$ .*

*Proof.* If  $(a, b) \in E(G(n))$  then  $(2^k a, 2^k b)$  is an edge in  $G(2^k n)$ , so we know that  $S = \{(2^k a, 2^k b) \mid (a, b) \in E(G(n))\}$  is a subset of  $E(G(2^k n))$ . The edges in  $S$  form a set of cycles which are isomorphic to  $G(n)$ . Hence, for all  $2^k v \in V(G(2^k n))$ ,  $2^k v$



**Figure 7.**  $G(28)$ .

is part of a cycle, so  $\text{indeg}(2^k v) \geq 1$ . Then by Theorem 6,  $\text{indeg}(2^k v) = 2$ . Thus,  $G(2^k n)$  contains a tree whose root vertex is  $2^k v$  for every  $v \in V(G(n))$ .

We now need to show that each of these trees is isomorphic to  $G(2^k)$  without the loop  $(0, 0)$ . Define  $T_v(2^k n)$  to be the tree whose root is  $r = 2^k v$ . Adapted from Theorem 5, each tree needs to satisfy the following three properties:

- (i)  $T_v(2^k n)$  has  $k + 1$  levels.
- (ii)  $T_v(2^k n)$  is a binary tree with exactly one vertex adjacent to  $r$  and  $\text{indeg}(v) = 0$  or  $2$  for all  $v \neq r$ .
- (iii) For any vertex  $v$  at level 0, the shortest path from  $v$  to  $r$  has length  $k$ .

First, Equation (4), we know that if  $a$  is the cyclical vertex adjacent to the root  $r = 2^k m$ , then  $s = a + 2^{k-1} n$  is also adjacent to  $r$  and  $2^{k-1} \parallel s$ . Thus, we have two vertices adjacent to  $r$ , and by Theorem 6,  $s$  is the only vertex in  $T_m(2^k n)$  that is adjacent to  $r$ . Thus, exactly one vertex in the tree is adjacent to  $r$ . The rest of part (ii) follows by definition from Theorem 6, so  $T_m(2^k n)$  is a binary tree and  $\text{indeg}(v) = 0$  or  $2$  for all  $v \neq r$ .

Now, for part (i), for any  $v \in V(T_m(2^k n))$  such that  $v \neq r$ , there exists an integer  $j \geq 0$  such that  $10^j v \equiv s = 2^{k-1} q \pmod{2^k n}$  for some integer  $q$  such that  $2 \nmid q$ . Suppose  $j > k - 1$ , so:

$$\begin{aligned} 10^j v - 2^{k-1} q &= 2^k n p \\ 2^{j-k+1} 5^j v - q &= 2 n p. \end{aligned}$$

This says that 2 divides  $2^{j-k+1} 5^j v - q$ . However,  $q$  is odd, so  $2^{j-k+1} 5^j v - q$  cannot be divisible by 2. Thus,  $j \leq k - 1$ .

Now assume  $j < k - 1$  for all  $v \in V(T_m(2^k n))$ . Then,

$$\begin{aligned} 10^j v - 2^{k-1} q &= 2^k np \\ 2^j 5^j v &= 2^k np + 2^{k-1} q \\ 5^j v &= 2^{k-1-j} (2np + q). \end{aligned} \tag{7}$$

This means that  $2 \mid v$  for all  $v \in V(T_m(2^k n))$ . From Theorem 6, all vertices in the tree now have an indegree of 2, which cannot be true as this would mean there are no vertex with an indegree of 0 and would make the graph an infinite tree. Thus, there exist vertices in  $T_m(2^k n)$  such that  $10^{k-1} v \equiv s$ , or such that the path from  $v$  to  $s$  is  $k - 1$  edges long, and hence the path from  $v$  to  $r$  is  $k$  edges long. Thus,  $T_m(2^k n)$  has  $k + 1$  levels.

Finally, from (7), we know that if the shortest path from  $v$  to  $s$  has length less than  $k - 1$ , then  $v$  must be even. Since all vertices at level 0 are odd, the shortest path from  $v$  at level 0 to  $s$  is  $k - 1$ , and the shortest path from level 0 to  $r$  has length  $k$ .

Therefore,  $T_v(2^k n)$  is isomorphic to the subgraph of  $G(2^k)$  without the loop  $(0, 0)$ . The root of each tree is  $2^k v$ , where  $v \in V(G(n))$ , so there are  $n$  of these trees.  $\square$

Theorem 7 is illustrated in  $G(28)$  (Figure 7) which contains 7 subgraphs isomorphic to  $G(4)$ . From this theorem, we also know that  $C_{2^k n} = C_n$  and  $L_{2^k n} = L_n$ .

Theorems 5 and 7 depended on the fact that 2 is a factor of 10. Thus, we can prove similar theorems for  $G(5^k)$  and  $G(5^k n)$  as well. From these, we can likewise determine that  $C_{5^k n} = C_n$  and  $L_{5^k n} = L_n$ .

## 5. Conclusion

The function  $f(x) = 10x \bmod n$  generates iteration digraphs whose cycles are greatly determined by the divisibility properties of  $n$ . With isomorphisms between  $G(n)$  and  $G(d)$ ,  $C_n$  is determined for any  $n$  relatively prime to 10. Then, 2 and 3 have specific relations to 10 which allow for simpler calculations for  $C_{2^k n}$  and  $C_{3n}$ . Thus, we can now calculate the number and lengths of cycles in  $G(n)$  for most integers  $n$ .

## References

- [Rosen 2000] K. H. Rosen, *Elementary number theory and its applications*, 4th ed., Addison-Wesley, Reading, MA, 2000. MR 2000i:11001 Zbl 0964.11002
- [Somer and Křížek 2004] L. Somer and M. Křížek, "On a connection of number theory with graph theory", *Czechoslovak Math. J.* **54(129)**:2 (2004), 465–485. MR 2005b:05112 Zbl 1080.11004
- [Wilson 2009] D. Wilson, "Divisibility by 7 is a walk on a graph", 2009, available at <http://blog.tanyakhovanova.com/?p=159>.

Received: 2012-11-05    Revised: 2013-03-04    Accepted: 2013-03-09

hjroberts3141@gmail.com    *College of Wooster, 91 Benton Street,  
Austintown, OH 44515, United States*

# Numerical integration of rational bubble functions with multiple singularities

Michael Schneier

(Communicated by Kenneth S. Berenhaut)

We derive an effective quadrature scheme via a partitioned Duffy transformation for a class of Zienkiewicz-like rational bubble functions proposed by J. Guzmán and M. Neilan. This includes a detailed construction of the new quadrature scheme, followed by a proof of exponential error convergence. Briefly discussed is the functions application to the finite element method when used to solve Stokes flow and elasticity problems. Numerical experiments which support the theoretical results are also provided.

## 1. Introduction

The finite element method is one of the most popular and well studied numerical methods used to approximate solutions of partial differential equations (PDEs). Its formulation is built upon the variational formulation of the PDE, where the infinite-dimensional problem is restricted to a finite-dimensional setting. What distinguishes the finite element method from other Galerkin methods is that the finite-dimensional space contains piecewise polynomials with respect to a partition (usually rectangles or triangles in two dimensions) of the domain. When performing the finite element method the need to integrate these piecewise polynomials over the partition arises. Solving these integrals directly would prove computationally costly and sometimes extremely difficult. We instead use a variety of numerical integration techniques. One of the most popular of these techniques is Gaussian quadrature. The method approximates the value of the integral via a weighted sum of function values at points within the domain of integration. This is already a mature theory for polynomial basis functions with highly developed implementation techniques and error analysis [Brezzi and Fortin 1991].

J. Guzmán and M. Neilan [2014a; 2014b] proposed a new family of finite methods to approximate two-dimensional Stokes flow and planar elasticity. Varying from the traditional finite element framework, the authors supplemented the usual finite

---

*MSC2010:* 65B99.

*Keywords:* Gaussian quadrature, multiple singularities, modified Duffy transformation.

element spaces (i.e., piecewise polynomials) with a class of divergence-free rational bubble functions. With the inclusion of these rational functions, Guzmán and Neilan were able to derive finite element methods with several desirable properties (e.g., exactly divergence-free velocity approximations for Stokes and symmetric and conforming stresses for elasticity). Assuming that the integrals are computed exactly, the authors derived several results including stability estimates of the numerical methods and optimal order error estimates. However in practice, these integrals are not computed exactly, and it is not clear how numerical integration will effect these theoretical results. The issue arises from the fact that traditional quadrature rules utilize interpolating polynomials to approximate the function and Taylor's formula to estimate the error [Burden and Faries 2011]. Thus in order to obtain accurate error estimates, our function must be sufficiently smooth. However, the rational functions in [Guzmán and Neilan 2014a; 2014b] are singular. Therefore the behavior of the error is unpredictable. We numerically verify this assertion in Section 5.

One of the traditional methods for computing the integrals of singular functions is the Duffy transformation [1982]. As described in [Lyness and Cools 1994], a mapping from the original triangular domain to the unit square is constructed. The singularity is effectively "stretched" out via its mapping to one of the edges of the square. Since the singularity is no longer present, the square can then be numerically integrated via a standard quadrature rule. This method will not work though for the divergence-free rational bubble functions described in this paper due to the presence of two singularities. While it would effectively eliminate one of the singularities, the remaining singularity would still render standard quadrature methods ineffective.

In the paper, we tackle this issue with a modified application of the Duffy transformation. We subdivide the triangle into four subtriangles and then perform a Duffy transformation on each of these subtriangles, which can essentially remove all the problematic singularities. We can then construct a quadrature rule on the unit square, which can then be mapped back to and used on our original domain. We do not address the effect of the error estimates obtained from this new scheme on the finite element methods in [Guzmán and Neilan 2014a; 2014b] as it is beyond the scope of this paper.

The remainder of this paper is organized as follows. Section 2 contains some preliminaries and the function spaces in which the analysis will be performed. Well known results from vector calculus which are used extensively in the analysis are also provided. In Section 3, the procedure for the partitioned Duffy transformation is established. In Section 4, a quadrature scheme derived from the partitioned Duffy transformation is given. A proof of exponential error convergence for this quadrature scheme is also provided. In Section 5, we present numerical experiments on the unit triangle which support our findings.



### 2. Preliminaries

In this paper, standard space and norm notations are adopted. If  $G : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is a mapping with argument  $x \in \mathbb{R}^n$ , we denote by  $DG(x)$  the Jacobian, that is,

$$DG_{ij}(x) = \frac{\partial G_i}{\partial x_j}(x) \quad (i = 1, 2, \dots, m, j = 1, 2, \dots, n).$$

For a differentiable function  $g : \mathbb{R}^n \rightarrow \mathbb{R}$ , we denote by  $\nabla g : \mathbb{R}^n \rightarrow \mathbb{R}^n$  the gradient of  $g$  which is given by

$$\nabla g(x) = \frac{\partial g}{\partial x_1}(x)e_1 + \dots + \frac{\partial g}{\partial x_n}(x)e_n,$$

where  $e_i$  is the orthogonal unit column vector pointing in the coordinates direction  $x_i$ . We note that  $\nabla g = (Dg)^t$  is the transpose of  $Dg$ . The Hessian matrix of a twice differentiable function  $g$  is denoted by  $D^2g : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$  and is defined as

$$D^2g(x) = D\nabla g(x), \tag{2-1}$$

where the operator  $D$  in (2-1) is applied row-wise. Namely, the Hessian matrix is given by

$$(D^2g)_{ij}(x) = \frac{\partial^2 g}{\partial x_i \partial x_j}(x) \quad (i, j = 1, 2, \dots, n).$$

For an open and bounded set  $D$  with Lipschitz continuous boundary  $\partial D$ , we denote by  $L^p(D)$  ( $1 \leq p \leq \infty$ ) the complete normed linear space

$$L^p(D) := \{ \text{measurable functions } v : \int_D |v|^p dx < \infty \} \quad (1 \leq p < \infty),$$

$$L^\infty(D) := \{ \text{measurable functions } v : \text{ess sup}_D |v| < \infty \}.$$

The corresponding norms are then given by

$$\|v\|_{L^p(D)} := \left( \int_D |v|^p dx \right)^{1/p}, \quad \|v\|_{L^\infty(D)} := \text{ess sup}_D |v|.$$

The Sobolev spaces  $W^{k,p}(D)$  are defined as

$$W^{k,p}(D) = \{ u \in L^p(D) : D^\alpha u \in L^p(D) \text{ for all } |\alpha| \leq k \},$$

with norms

$$\|u\|_{W^{k,p}(D)} = \left( \sum_{|\alpha| \leq k} \int_D |D^\alpha u|^p dx \right)^{1/p} \quad (1 \leq p < \infty)$$

and

$$\|u\|_{W^{k,\infty}(D)} = \sum_{|\alpha|\leq k} \operatorname{ess\,sup}_D |D^\alpha u|.$$

In the case  $p = 2$  and  $k \geq 1$ , we set  $H^k(D) = W^{k,2}(D)$  and  $\|\cdot\|_{H^k(D)} = \|\cdot\|_{W^{k,2}(D)}$ . We note that  $H^k(D)$  is a Hilbert space.

We denote the dual space of  $W^{k,p}(D)$  by  $W^{-k,p'}(D)$ , where  $p'$  satisfies

$$\frac{1}{p} + \frac{1}{p'} = 1.$$

The associated norm is defined by

$$\|\varphi\|_{W^{-k,p'}(D)} = \sup_{v \in W^{k,p}(D) \setminus \{0\}} \varphi(v) / \|v\|_{W^{k,p}(D)}. \tag{2-2}$$

We denote by  $\mathcal{T}_h$  a shape-regular triangulation of the domain  $\Omega$  with  $h_T = \operatorname{diam}(T)$  for all  $T \in \mathcal{T}_h$  and  $h := \max_{T \in \mathcal{T}_h} h_T$ . Given  $T \in \mathcal{T}_h$ , we denote by  $\{e^{(i)}\}_{i=1}^3$  the three edges of  $T$  and by  $\{\lambda^{(i)}\}_{i=1}^3$  the three barycentric coordinates labeled such that  $\lambda^{(i)}|_{e^{(i)}} = 0$ . The vertices of  $T$  are denoted by  $\{a^{(i)}\}_{i=1}^3$  labeled such that  $\lambda^{(i)}(a^{(i)}) = \delta_{i,j}$ . We set  $b_T := \lambda^{(1)}\lambda^{(2)}\lambda^{(3)} \in \mathcal{P}_3(T)$  to be the cubic bubble and  $b^{(i)} = \lambda^{(i+1)}\lambda^{(i+2)} \in \mathcal{P}_2(T) \pmod{3}$  to be the quadratic edge bubble associated with edge  $e^{(i)}$ . For each triangle  $T \in \mathcal{T}_h$ , the three *rational edge bubbles*  $\{B^{(i)}\}_{i=1}^3$  associated with  $T$  are then given by

$$\begin{aligned} B^{(i)} &:= \frac{b_T b^{(i)}}{(\lambda^{(i)} + \lambda^{(i+1)})(\lambda^{(i)} + \lambda^{(i+2)})} && \text{if } 0 \leq \lambda^{(i)} \leq 1, 0 \leq \lambda^{(i+1)}, \lambda^{(i+2)} < 1, \\ B^{(i)}(a^{(i+1)}) = B^{(i)}(a^{(i+2)}) &= 0 && \text{otherwise.} \end{aligned} \tag{2-3}$$

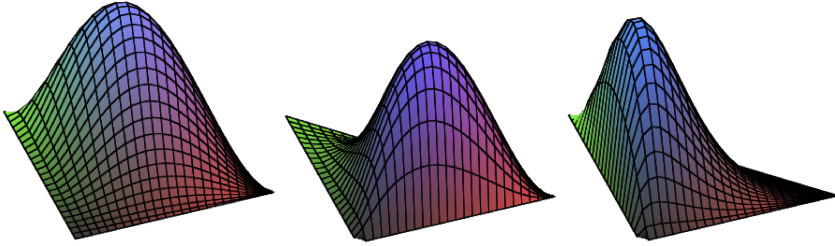
The graphs of the three rational bubble functions are depicted in Figure 1 on the reference triangle with vertices  $(0, 0)$ ,  $(0, 1)$  and  $(1, 0)$ . In this case, the barycentric coordinates reduce to  $\hat{\lambda}^{(1)} = x_1$ ,  $\hat{\lambda}^{(2)} = x_2$  and  $\hat{\lambda}^{(3)} = 1 - x_1 - x_2$ . Therefore, the three rational bubble functions on the reference triangle are given by

$$\begin{aligned} B^{(1)} &= \frac{x_1 x_2^2 (1 - x_1 - x_2)^2}{(x_1 + x_2)(1 - x_2)}, & B^{(2)} &= \frac{x_2 x_1^2 (1 - x_1 - x_2)^2}{(x_1 + x_2)(1 - x_1)}, \\ B^{(3)} &= \frac{x_1^2 x_2^2 (1 - x_1 - x_2)}{(1 - x_1)(1 - x_2)}. \end{aligned} \tag{2-4}$$

In [Guzmán and Neilan 2014a] (see also [Ciarlet 1978, pp. 347–348]), the following lemma pertaining to the rational bubble functions was established.

**Lemma 2.1.** *For each  $T \in \mathcal{T}_h$ , the following hold ( $i = 1, 2, 3$ ):*

$$B^{(i)} \in C^1(\bar{T}) \cap W^{2,\infty}(T), \quad B^{(i)}|_{\partial T} = 0, \quad \nabla B^{(i)}(a^{(j)}) = 0 \quad (j = 1, 2, 3). \tag{2-5}$$



**Figure 1.** The graphs of the three bubble functions on the reference triangle with vertices  $(0, 0)$  (left),  $(0, 1)$  (middle) and  $(1, 0)$  (right).

We end this section by stating some well known vector calculus results, which will be used extensively in the analysis below.

**Theorem 2.2** (inverse function theorem [Spivak 1998]). *Suppose that  $G : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is continuously differentiable in an open set containing a point  $a$  with  $\det(DG(a)) \neq 0$ . Then there is an open set  $V$  containing  $a$  and an open set  $W$  containing  $G(a)$  such that  $G : V \rightarrow W$  has a continuous inverse  $G^{-1} : W \rightarrow V$  which is differentiable for all  $y \in W$ . Moreover, there holds*

$$D(G^{-1})(y) = [DG(G^{-1}(y))]^{-1}. \tag{2-6}$$

**Lemma 2.3** (Bramble–Hilbert lemma [Ciarlet 1978]). *Let  $D$  be an open subset of  $\mathbb{R}^n$  ( $n \geq 1$ ) with a Lipschitz-continuous boundary. For some  $k \geq 0$  and some number  $p \in [0, \infty]$ , let  $\varphi$  be a continuous linear form on the space  $W^{k+1,p}(D)$  with the property that*

$$\varphi(p) = 0 \quad \text{for all } p \in P_k(D),$$

where  $P_k(D)$  is the set of all polynomials up to order  $k$  on  $D$ . Then there exists a positive constant  $C > 0$  depending on  $D$  such that for all  $v \in W^{k+1,p}(D)$ ,

$$|\varphi(v)| \leq C \|\varphi\|_{W^{-k-1,p'}(D)} |v|_{W^{k+1,p}(D)},$$

where  $\|\cdot\|_{W^{-k-1,p'}(D)}$  is defined by (2-2).

**Theorem 2.4** (Sard’s theorem [Spivak 1998]). *Let  $G : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  be a continuously differentiable mapping. Let  $X$  be the set of points  $x$  in  $\mathbb{R}^2$  at which the Jacobian matrix  $DG(x)$  has rank less than 2. Then  $G(X)$  has Lebesgue measure 0 in  $\mathbb{R}^2$ .*

**Corollary 2.5.** *Let  $V, U \subset \mathbb{R}^2$  be two open and bounded sets, and let  $G : V \rightarrow U$  be a continuously differentiable mapping that is surjective onto  $U$ ; that is,  $G(V) = U$ .*

Define the set  $X = \{x \in V : DG(x) \text{ does not have full rank}\}$ . Then for any continuous function  $f \in C^0(U)$ ,

$$\int_U f(y) dy = \int_{G(V) \setminus G(X)} f(y) dy.$$

*Proof.* Since  $U = G(V)$ , we have

$$\int_U f(y) dy = \int_{G(V)} f(y) dy = \int_{G(V) \setminus G(X)} f(y) dy + \int_{G(X)} f(y) dy.$$

From Sard's lemma,  $G(x)$  has Lebesgue measure 0. Therefore  $\int_{G(X)} f(y) dy = 0$ , and so

$$\int_U f(y) dy = \int_{G(V)} f(y) dy = \int_{G(V) \setminus G(X)} f(y) dy. \quad \square$$

### 3. A partitioned Duffy transform

In this section, we describe a partitioned Duffy transform which essentially removes the singularities of the rational bubble functions defined by (2-3). Basically the strategy is to subdivide each triangle into four subtriangles by a red refinement and then apply the Duffy transform to each subtriangle that shares a vertex with the parent triangle. To describe this procedure in further detail, we require some notation.

Denote by  $\hat{T}$  the unit triangle with vertices  $\hat{a}^{(1)} := (1, 0)$ ,  $\hat{a}^{(2)} := (0, 1)$  and  $\hat{a}^{(3)} := (0, 0)$ , and let  $\{\hat{K}^{(i)}\}_{i=1}^4$  be the four subtriangles of  $\hat{T}$  obtained by connecting the three midpoints of each edge of  $\hat{T}$  (see Figure 2), where  $\hat{a}^{(i)}$  is a vertex of  $\hat{K}^{(i)}$  ( $i = 1, 2, 3$ ). We denote the three vertices of  $\hat{K}^{(i)}$  by  $\{\hat{b}_j^{(i)}\}_{j=1}^3$  oriented in a counterclockwise fashion and labeled such that

$$\hat{a}^{(i)} = \hat{b}_i^{(i)} \quad (i = 1, 2, 3).$$

The vertices  $\{\hat{b}_j^{(4)}\}_{j=1}^3$  can be labeled arbitrarily. Define  $\hat{F}_i : \hat{T} \rightarrow \hat{K}^{(i)}$  to be the affine mapping such that  $\hat{F}_i(\hat{a}^{(i)}) = \hat{b}^{(i)}$  ( $i = 1, 2, 3$ ); that is,

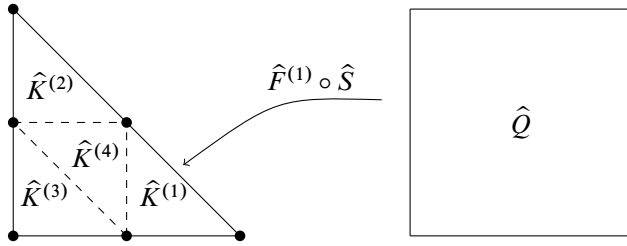
$$\hat{F}_1(y) = \left(\frac{1}{2}(y_1 + 1), \frac{1}{2}y_2\right), \tag{3-1}$$

$$\hat{F}_2(y) = \left(\frac{1}{2}(1 - y_1 - y_2), \frac{1}{2}(y_1 + 1)\right), \tag{3-2}$$

$$\hat{F}_3(y) = \left(\frac{1}{2}y_2, -\frac{1}{2}(y_1 + y_2 - 1)\right). \tag{3-3}$$

In the case  $i = 4$ ,  $\hat{F}_i$  can be any one of the possible affine mappings that takes  $\hat{T}$  onto  $\hat{K}^{(4)}$ . Denote by  $\hat{Q} := (0, 1)^2$  the unit square and define the Duffy transform  $\hat{S} : \hat{Q} \rightarrow \hat{T}$  as

$$\hat{S}(\hat{s}) := (\hat{s}_1, \hat{s}_2(1 - \hat{s}_1))^t. \tag{3-4}$$



**Figure 2.** A pictorial description of the notation used.

Finally for a function  $f : \hat{T} \rightarrow \mathbb{R}$ , we set

$$\hat{f}_i(\hat{s}) := f(\hat{F}_i(\hat{S}(\hat{s}))) \quad (i = 1, 2, 3) \quad \text{and} \quad \hat{f}_4(\hat{y}) = f(F_4(\hat{y})). \quad (3-5)$$

We note that  $\hat{f}_i : \hat{Q} \rightarrow \mathbb{R} \quad (i = 1, 2, 3)$ , whereas  $\hat{f}_4 : \hat{T} \rightarrow \mathbb{R}$ .

**Lemma 3.1.** For  $i = 1, 2, 3$ , define  $\hat{G}_i = F_i \circ \hat{S} : \hat{Q} \rightarrow \hat{K}^{(i)}$ . Then there holds

$$(\widehat{\nabla_{\hat{x}} f})_i(\hat{s}) = (D_{\hat{s}} G_i(\hat{s}))^{-t} \nabla_{\hat{s}} \hat{f}_i(\hat{s}),$$

for any function  $f$  satisfying  $\hat{f}_i \in C^1(\hat{Q})$ . Here,  $\nabla_{\hat{x}}$  and  $\nabla_{\hat{s}}$  denotes the gradient with respect to  $\hat{x}$  and  $\hat{s}$ , respectively  $D_{\hat{s}} = (\nabla_{\hat{s}})^t$ , and  $(D_{\hat{s}} G_i(\hat{s}))^{-t}$  denote the inverse matrix of the transpose of  $D_{\hat{s}} G_i(\hat{s})$ .

*Proof.* For ease of notation, we omit the subscript  $i$  in the arguments below.

By (3-5) and the definition of  $\hat{G}$ , we have  $\hat{f}(\hat{s}) = f(\hat{G}(\hat{s}))$ . Now let  $\hat{x} = G(\hat{s})$  so that  $\hat{f}(\hat{s}) = f(\hat{x})$  and  $\hat{s} = G^{-1}(\hat{x})$ . We then have

$$\hat{s}_k = (G^{-1})_k(\hat{x}) \quad \text{and} \quad \frac{\partial \hat{s}_k}{\partial \hat{x}_j} = \frac{\partial}{\partial \hat{x}_j} (G^{-1})_k(\hat{x}).$$

Letting  $D_{\hat{x}} G^{-1}(\hat{x})$  be the Jacobian of  $G^{-1}$ , we have

$$\frac{\partial \hat{s}_k}{\partial \hat{x}_j} = (D G^{-1})_{kj}(\hat{x}). \quad (3-6)$$

Therefore by the chain rule and (3-6), we have

$$\begin{aligned} \left( \frac{\partial f}{\partial \hat{x}_j} \circ G \right)(\hat{s}) &= \sum_{k=1}^2 \frac{\partial f}{\partial \hat{s}_k}(\hat{s}) \frac{\partial \hat{s}_k}{\partial \hat{x}_j} = \sum_{k=1}^2 \frac{\partial f}{\partial \hat{s}_k}(\hat{s}) (D_{\hat{x}} G^{-1}(\hat{x}))_{kj} \\ &= \sum_{k=1}^2 ((D_{\hat{x}} G^{-1}(\hat{x}))_{jk})^t \frac{\partial f}{\partial \hat{s}_k}(\hat{s}) = ((D_{\hat{x}} G^{-1})^t(\hat{x}) \nabla_{\hat{s}} \hat{f}(\hat{s}))_j. \end{aligned}$$

It then follows that

$$(\nabla_x f \circ G)(\hat{s}) = (D_{\hat{x}} G^{-1}(\hat{x}))^t \nabla_{\hat{s}} \hat{f}(\hat{s}). \quad (3-7)$$

Now by the implicit function theorem (see Theorem 2.2), we have

$$D_{\hat{x}} G^{-1}(\hat{x}) = [DG(G^{-1}(\hat{x}))]^{-1} = [DG(\hat{s})]^{-1}.$$

Therefore by (3-7) and (3-5), we have

$$\widehat{\nabla_{\hat{x}} f}(\hat{s}) = (\nabla_{\hat{x}} f \circ G)(\hat{s}) = (D_{\hat{s}} G(\hat{s}))^{-t} \nabla_{\hat{s}} \hat{f}(\hat{s}). \quad \square$$

**Lemma 3.2.** *Let  $\widehat{G}_i = \widehat{F}_i \circ \widehat{S} : \widehat{Q} \rightarrow \widehat{K}^{(i)}$ . Then,*

$$(\widehat{D_{\hat{x}}^2 f})_i(\hat{s}) = D_{\hat{s}}((D_{\hat{s}} G_i(\hat{s}))^{-t} \nabla_{\hat{s}} \hat{f}_i(\hat{s})) D_{\hat{s}} G_i(\hat{s})^{-1}.$$

*Proof.* Again, we omit the subscript  $i$  in the proof for ease of notation.

From Lemma 3.1 we have

$$\left( \frac{\partial f}{\partial \hat{x}_j} \circ G \right)(\hat{s}) = \sum_{k=1}^2 (D_{\hat{s}} G(\hat{s}))_{jk}^{-t} \frac{\partial \hat{f}}{\partial \hat{s}_k}(\hat{s}).$$

Set

$$r_{jk}(\hat{s}) := (D_{\hat{s}} G(\hat{s}))_{jk}^{-t} \frac{\partial \hat{f}}{\partial \hat{s}_k}(\hat{s}) \quad \text{so that} \quad \frac{\partial f}{\partial \hat{x}_j}(\hat{x}) = \sum_{k=1}^2 r_{jk}(\hat{s}).$$

Then by the chain rule, (3-6), and the inverse function theorem, we have

$$\begin{aligned} \left( \frac{\partial^2 f}{\partial \hat{x}_j \partial \hat{x}_l} \circ G \right)(\hat{s}) &= \sum_{k=1}^2 \sum_{m=1}^2 \frac{\partial r_{jk}}{\partial \hat{s}_m}(\hat{s}) \frac{\partial \hat{s}_m}{\partial \hat{x}_l} = \sum_{k=1}^2 \sum_{m=1}^2 \frac{\partial r_{jk}}{\partial \hat{s}_m}(\hat{s}) (D_{\hat{x}} G^{-1})_{ml}(\hat{x}) \\ &= \sum_{k=1}^2 \sum_{m=1}^2 \frac{\partial r_{jk}}{\partial \hat{s}_m}(\hat{s}) (D_{\hat{x}} G)_{ml}^{-1}(\hat{s}). \end{aligned}$$

Now since

$$r_{jk}(\hat{s}) = (D_{\hat{s}} G(\hat{s}))_{jk}^{-t} \frac{\partial \hat{f}}{\partial \hat{s}_k}(\hat{s}),$$

we have

$$\begin{aligned} \left(\frac{\partial^2 f}{\partial \hat{x}_j \partial \hat{x}_l} \circ G\right)(\hat{s}) &= \sum_{k=1}^2 \sum_{m=1}^2 \frac{\partial}{\partial \hat{s}_m} \left( (D_{\hat{s}} G)_{jk}^{-t}(\hat{s}) \frac{\partial \hat{f}}{\partial \hat{s}_k}(\hat{s}) \right) ((D_{\hat{s}} G)_{ml}^{-1}(\hat{s})) \\ &= \sum_{m=1}^2 \frac{\partial}{\partial \hat{s}_m} \left( (D_{\hat{s}} G)^{-t}(\hat{s}) \nabla_{\hat{s}} \hat{f}(\hat{s}) \right)_j ((D_{\hat{s}} G)_{ml}^{-1}(\hat{s})) \\ &= \sum_{m=1}^2 (D_{\hat{s}} ((D_{\hat{s}} G)^{-t}(\hat{s}) \nabla_{\hat{s}} \hat{f}(\hat{s})))_{jm} ((D_{\hat{s}} G)_{ml}^{-1}(\hat{s})) \\ &= \left( (D_{\hat{s}} ((D_{\hat{s}} G)^{-t}(\hat{s}) \nabla_{\hat{s}} \hat{f}(\hat{s}))) ((D_{\hat{s}} G)^{-1}(\hat{s})) \right)_{jl}. \end{aligned}$$

It then follows that

$$\widehat{(D_{\hat{x}}^2 f)}_i(\hat{s}) = (D_{\hat{x}}^2 f \circ \hat{G}_i)(\hat{s}) = D_{\hat{s}} \left( (D_{\hat{s}} G_i(\hat{s}))^{-t} \nabla_{\hat{s}} \hat{f}_i(\hat{s}) \right) D_{\hat{s}} G_i(\hat{s})^{-1}. \quad \square$$

We are now ready to state the main result of this section.

**Lemma 3.3.** *Let  $B^{(j)}$  be the rational edge bubble (2-3) defined on the reference triangle  $\hat{T}$  with vertices  $(1, 0)$ ,  $(0, 1)$ , and  $(0, 0)$ . Then  $(i = 1, 2, 3)$ ,*

$$\widehat{B^{(j)}}_i \in C^\infty(\hat{Q}), \quad (\nabla_{\hat{x}} \widehat{B^{(j)}})_i \in [C^\infty(\hat{Q})]^2, \quad (D_{\hat{x}}^2 \widehat{B^{(j)}})_i \in [C^\infty(\hat{Q})]^{2 \times 2},$$

and

$$\widehat{B^{(j)}}_4 \in C^\infty(\hat{T}), \quad (\nabla_{\hat{x}} \widehat{B^{(j)}})_4 \in [C^\infty(\hat{T})]^2, \quad (D_{\hat{x}}^2 \widehat{B^{(j)}})_4 \in [C^\infty(\hat{T})]^{2 \times 2}.$$

Here,  $\nabla_{\hat{x}}$  denotes the gradient with respect to  $\hat{x}$  and  $D_{\hat{x}}^2$  denotes the Hessian with respect to  $\hat{x}$ .

**Remark 3.4.** Essentially, Lemma 3.3 states that if we map the rational bubble functions' derivatives to the unit square via the partitioned Duffy transform, then the resulting function is  $C^\infty$ .

*Proof.* Due to the symmetry of the rational edge bubbles, it suffices to prove the result for the function

$$B(\hat{x}) := \frac{\hat{x}_1^2 \hat{x}_2^2 (1 - \hat{x}_1 - \hat{x}_2)}{(1 - \hat{x}_1)(1 - \hat{x}_2)} \in C^1(\hat{T}) \cap W^{2,\infty}(\hat{T}) \tag{3-8}$$

(see (2-4)). Since  $B(\hat{x})$  has singularities only at the vertices  $(1, 0)$  and  $(0, 1)$ , we have  $B|_{K_4} \in C^\infty(\hat{K}_4)$ . It is then trivial to see that

$$\widehat{B}_4 \in C^\infty(\hat{T}), \quad (\nabla_{\hat{x}} \widehat{B})_4 \in [C^\infty(\hat{T})]^2, \quad (D_{\hat{x}}^2 \widehat{B})_4 \in [C^\infty(\hat{T})]^{2 \times 2}.$$

Next, a direct calculation shows that

$$\begin{aligned}\widehat{B}_1(\hat{s}) &= B(F_1(S(\hat{s}))) = B\left(\frac{1}{2}(\hat{s}_1 + 1), \frac{1}{2}(\hat{s}_2(1 - \hat{s}_1))\right) \\ &= \frac{\hat{s}_2^2(\hat{s}_2 - 1)(\hat{s}_1 - 1)^2(\hat{s}_1 + 1)^2}{8(2 - \hat{s}_2 + \hat{s}_2\hat{s}_1)} \in C^\infty(\widehat{Q}),\end{aligned}\quad (3-9a)$$

$$\begin{aligned}\widehat{B}_2(\hat{s}) &= B(F_2(S(\hat{s}))) = B\left(\frac{1}{2}(1 - \hat{s}_1 - \hat{s}_2(1 - \hat{s}_1)), \frac{1}{2}(\hat{s}_1 + 1)\right) \\ &= \frac{1}{8} \frac{(s_1 - 1)s_2(s_1 + 1)^2(s_2 - 1)(1 - s_1 - s_2 + s_2s_1)}{8(1 + s_1 + s_2 - s_2s_1)} \in C^\infty(\widehat{Q}),\end{aligned}\quad (3-9b)$$

$$\begin{aligned}\widehat{B}_3(\hat{s}) &= B(F_3(S(\hat{s}))) = B\left(\frac{1}{2}\hat{s}_2(1 - \hat{s}_1), -\frac{1}{2}(\hat{s}_1 + \hat{s}_2(1 - \hat{s}_1)) + \frac{1}{2}\right) \\ &= \frac{\hat{s}_2^2(\hat{s}_1 - 1)^2(1 - \hat{s}_1 - \hat{s}_2 + \hat{s}_2\hat{s}_1)^2(1 + \hat{s}_1)}{8(2 - \hat{s}_2 + \hat{s}_2\hat{s}_1)(1 + \hat{s}_1 + \hat{s}_2 - \hat{s}_2\hat{s}_1)} \in C^\infty(\widehat{Q}).\end{aligned}\quad (3-9c)$$

Then from Lemma 3.1, we have

$$(\widehat{\nabla_{\hat{x}} B})_i(\hat{s}) = (D_{\hat{s}} G_i(\hat{s}))^{-t} \nabla_{\hat{s}} \widehat{B}(\hat{s}), \quad (3-10)$$

where  $\nabla_{\hat{s}}$  denotes the gradient with respect to  $\hat{s}$ ,  $D_{\hat{s}} = (\nabla_{\hat{s}})^t$ , and  $(D_{\hat{s}} G_i(\hat{s}))^{-t}$  denotes the inverse matrix of the transpose of  $D_{\hat{s}} G_i(\hat{s})$ . Using the identity  $DF_i = 2|K_i| = 1/2$  and the chain rule, we have

$$DG_i(\hat{s}) = D(F_i(S(\hat{s}))) = DF_i(S(\hat{s}))DS(\hat{s}) = \frac{1}{2}DS(\hat{s}).$$

It then follows that  $(DG_i(\hat{s}))^{-t} = 2(DS(\hat{s}))^{-t}$ ; that is,

$$(DG_i(\hat{s}))^{-t} = \begin{pmatrix} 2 & 2\hat{s}_2/(1 - \hat{s}_1) \\ 0 & 2/(1 - \hat{s}_1) \end{pmatrix}. \quad (3-11)$$

By (3-9), we see that the derivatives  $\partial \widehat{B}_i / \partial \hat{s}_2$  ( $i = 1, 2, 3$ ) all have a factor  $(1 - \hat{s}_1)^2$ . In particular, we may write

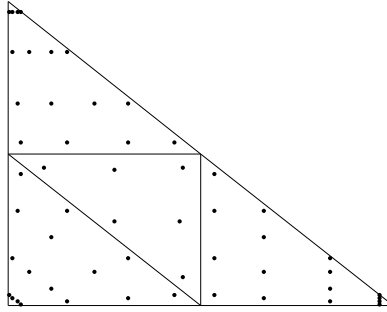
$$\nabla_{\hat{s}} \widehat{B}_i = \begin{pmatrix} g_i^{(1)}(\hat{s})(1 - \hat{s}_1) \\ g_i^{(2)}(\hat{s})(1 - \hat{s}_1)^2 \end{pmatrix} \quad (3-12)$$

for some  $g_i^{(1)}, g_i^{(2)} \in C^\infty(\widehat{Q})$ . Combining (3-12) with (3-10) and (3-11) we see that  $(\widehat{\nabla_{\hat{x}} B})_i \in [C^\infty(\widehat{Q})]^2$ .

Continuing, we use Lemma 3.2 and the inverse function theorem to obtain

$$(\widehat{D_{\hat{x}}^2 B})_i(\hat{s}) = D_{\hat{s}}((D_{\hat{s}} G_i(\hat{s}))^{-t} \nabla_{\hat{s}} \widehat{B}(\hat{s})) D_{\hat{s}} G_i(\hat{s})^{-1}. \quad (3-13)$$





**Figure 3.** The location of the nodes  $\hat{x}^{(j)}$  with  $L = 16$  and  $M = 6$ .

By (3-12) and (3-11), we have

$$D_{\hat{s}}((D_{\hat{s}}G_i(\hat{s}))^{-t}\nabla_{\hat{s}}\widehat{B}(\hat{s})) = \begin{pmatrix} g_i^{(1,1)}(\hat{s}) & g_i^{(1,2)}(\hat{s})(1-\hat{s}_1) \\ g_i^{(2,1)}(\hat{s}) & g_i^{(2,2)}(\hat{s})(1-\hat{s}_1) \end{pmatrix},$$

for some  $g^{(i,j)}(\hat{s}) \in C^\infty(\widehat{Q})$ . It then follows from the definition of  $D_{\hat{s}}G_i(\hat{s})^{-1}$  (see (3-11)) and (3-13) that

$$(\widehat{D_{\hat{x}}^2 B})_i(\hat{s}) \in [C^\infty(\widehat{Q})]^{2 \times 2}. \quad \square$$

**4. A quadrature rule based upon the partitioned Duffy transform**

We now build quadrature schemes for the integral  $\int_{\widehat{T}} f(\hat{x}) d\hat{x}$  based upon the partitioned Duffy transform described above. To this end, we let  $\{\hat{s}^{(j)}, \hat{\theta}^{(j)}\}_{j=1}^L$  be a tensor product Gaussian quadrature rule on the unit square  $\widehat{Q}$ , and we let  $\{\hat{y}^{(j)}, \hat{\varrho}^{(j)}\}_{j=1}^M$  be a quadrature rule on the unit triangle  $\widehat{T}$ . We then map the quadrature points and weights on  $\widehat{Q}$  to the subtriangles  $\widehat{K}^{(i)}$  ( $i = 1, 2, 3$ ) by the formulas  $\hat{x}^{((i-1)L+j)} = \widehat{F}_i(\widehat{S}(\hat{s}^{(j)}))$  and

$$\hat{\omega}^{((i-1)L+j)} = \frac{1}{2}(1-\hat{s}_1^{(j)})\hat{\theta}^{(j)} \quad (j = 1, 2, \dots, L).$$

We map the quadrature points and weights  $\{\hat{y}^{(j)}, \hat{\varrho}^{(j)}\}_{j=1}^M$  to  $\widehat{K}^{(4)}$  by

$$\hat{x}^{3L+j} = \widehat{F}_4(\hat{y}^{(j)}) \quad \text{and} \quad \hat{\omega}^{(3L+j)} = \frac{1}{2}\hat{\varrho}^{(j)} \quad (j = 1, 2, \dots, M).$$

The new quadrature scheme on  $\widehat{T}$  is then given by  $\{\hat{x}^{(j)}, \hat{\omega}^{(j)}\}_{j=1}^{3L+M}$  (see Figure 3).

**Remark 4.1.** By Figure 3, we see that the quadrature points are clustered near the vertices of the (macro) triangle  $\widehat{T}$ . On the other hand, the weights defined by  $\hat{\omega}^{((i-1)L+j)} = \frac{1}{2}(1-\hat{s}_1^{(j)})\hat{\theta}^{(j)}$  are small near the vertices since the line  $\hat{s}_1 = 1$  on  $\widehat{Q}$  is mapped to each of the vertices of  $\widehat{T}$ .

Note by a change of variables and Sard's theorem, we have

$$\begin{aligned}
 \int_{\hat{T}} f(\hat{x}) d\hat{x} &= \sum_{i=1}^4 \int_{\hat{K}^{(i)}} f(\hat{x}) d\hat{x} = \sum_{i=1}^4 \int_{\hat{T}} f(\hat{F}^{(i)}(\hat{y})) |D\hat{F}^{(i)}(\hat{y})| d\hat{y} \\
 &= \sum_{i=1}^4 2|\hat{K}^{(i)}| \int_{\hat{T}} f(\hat{F}_i(\hat{y})) d\hat{y} \\
 &= \frac{1}{2} \sum_{i=1}^3 \int_{\hat{T}} f(\hat{F}_i(\hat{y})) d\hat{y} + \frac{1}{2} \int_{\hat{T}} \hat{f}_4(\hat{y}) d\hat{y} \\
 &= \frac{1}{2} \sum_{i=1}^3 \int_{\hat{Q}} f(\hat{F}_i(\hat{S}(\hat{s}))) |D_{\hat{s}} \hat{S}(\hat{s})| d\hat{s} + \frac{1}{2} \int_{\hat{T}} \hat{f}_4(\hat{y}) d\hat{y} \\
 &= \frac{1}{2} \sum_{i=1}^3 \int_{\hat{Q}} \hat{f}_i(\hat{s})(1 - \hat{s}_1) d\hat{s} + \frac{1}{2} \int_{\hat{T}} \hat{f}_4(\hat{y}) d\hat{y}.
 \end{aligned}$$

We also have

$$\begin{aligned}
 \sum_{j=1}^{3L+M} \hat{\omega}^{(j)} f(\hat{x}^{(j)}) &= \sum_{i=1}^3 \sum_{j=1}^L \hat{\omega}^{(i-1)L+j} f(\hat{x}^{((i-1)L+j)}) + \sum_{j=1}^M \hat{\omega}^{(3L+j)} f(\hat{x}^{(3L+j)}) \\
 &= \frac{1}{2} \sum_{i=1}^3 \sum_{j=1}^L (1 - \hat{s}^{(j)}) \theta^{(j)} f(\hat{F}_i(\hat{S}(\hat{s}^{(j)}))) + \frac{1}{2} \sum_{j=1}^M \hat{\varrho}^{(j)} f(\hat{F}_4(\hat{y}^{(j)})) \\
 &= \frac{1}{2} \sum_{i=1}^3 \sum_{j=1}^L \theta^{(j)} \hat{f}_i(\hat{s}^{(j)})(1 - \hat{s}_1^{(j)}) + \frac{1}{2} \sum_{j=1}^M \hat{\varrho}^{(j)} \hat{f}_4(\hat{y}^{(j)}).
 \end{aligned}$$

It then follows from these two identities that the error can be written as

$$\begin{aligned}
 E_{\hat{T}}(f) &:= \int_{\hat{T}} f(\hat{x}) d\hat{x} - \sum_{j=1}^{3L+M} \hat{\omega}^{(j)} f(\hat{x}^{(j)}) \tag{4-1} \\
 &= \frac{1}{2} \sum_{i=1}^3 \left( \int_{\hat{Q}} \hat{f}_i(\hat{s})(1 - \hat{s}_1) d\hat{s} - \sum_{j=1}^L \theta^{(j)} \hat{f}_i(\hat{s}^{(j)})(1 - \hat{s}_1^{(j)}) \right) \\
 &\quad + \frac{1}{2} \left( \int_{\hat{T}} \hat{f}_4(\hat{y}) d\hat{y} - \sum_{j=1}^M \hat{\varrho}^{(j)} \hat{f}_4(\hat{y}^{(j)}) \right).
 \end{aligned}$$

**Theorem 4.2.** *Let  $B^{(j)}$  be any of the three rational edge bubbles defined on the reference triangle. Then for any multi-index  $\alpha = (\alpha_1, \alpha_2)$  with  $0 \leq |\alpha| \leq 2$ , there exists  $C_\alpha > 0$  and  $\delta_\alpha > 0$  such that*

$$\left| E_{\hat{T}} \left( \frac{\partial^{|\alpha|} B^{(j)}}{\partial \hat{x}^\alpha} \right) \right| \leq C_\alpha (\exp(-\delta_\alpha M) + \exp(-\delta_\alpha L)).$$

*Proof.* This result follows from (4-1), Lemma 3.3, and standard estimates of Gaussian quadrature [Sauter and Schwab 2011, pp. 324–325].  $\square$

We now discuss the quadrature rule on an arbitrary triangle  $T \in \mathcal{T}_h$ . This is done in a natural way. Namely, letting  $F_T : \hat{T} \rightarrow T$  denote the affine transformation, we define the quadrature scheme  $\{x_T^{(j)}, \omega_T^{(j)}\}_{j=1}^{3L+M}$  by  $x_T^{(j)} = F_T(\hat{x}^{(j)})$  and  $\omega_T^{(j)} = 2|T|\hat{\omega}^{(j)}$ . The error of the scheme is then given by

$$E_T(f) := \int_T f(x) dx - \sum_{j=1}^{3L+M} \omega_T^{(j)} f(x_T^{(j)}).$$

Using the Bramble–Hilbert lemma, we can obtain the following result.

**Theorem 4.3.** *Suppose that the quadrature schemes*

$$\{\hat{s}^{(j)}, \hat{\theta}^{(j)}\}_{j=1}^L \quad \text{and} \quad \{\hat{y}^{(j)}, \hat{q}^{(j)}\}_{j=1}^M$$

*are exact for polynomials of degree at most  $m$  on  $\hat{Q}$  and  $\hat{T}$ , respectively. For a given triangle  $T \in \mathcal{T}_h$ , let  $f$  be a continuous function on  $\bar{T}$  and  $\hat{f}_i \in H^{m+1}(\hat{Q})$  and  $\hat{f}_4 \in H^{m+1}(\hat{T})$ , where*

$$\hat{f}_i(\hat{s}) = f(F_T(\hat{F}_i(\hat{S}(\hat{s})))) \quad \text{and} \quad \hat{f}_4(\hat{y}) = f(F_T(\hat{F}_4(\hat{y}))).$$

*Then,*

$$E_T(f) \leq Ch_T^2 \left( \sum_{i=1}^3 |\hat{f}_i|_{H^{m+1}(\hat{Q})} + |\hat{f}_4|_{H^{m+1}(\hat{T})} \right).$$

*Proof.* Let  $\hat{f} \in C^0(\hat{T})$  be defined as  $\hat{f}(\hat{x}) = f(F_T(\hat{x}))$  so that  $\hat{f}_i(\hat{s}) = \hat{f}(\hat{F}_i(\hat{S}(\hat{s})))$  and  $\hat{f}_4(\hat{y}) = \hat{f}(\hat{F}_4(\hat{y}))$ . Then by a change of variables and (4-1), we have

$$\begin{aligned} E_T(f) &= 2|T|E_{\hat{T}}(\hat{f}) \\ &= |T| \left( \sum_{i=1}^3 \left( \int_{\hat{Q}} \hat{f}_i(\hat{s})(1-\hat{s}_1) d\hat{s} - \sum_{j=1}^L \hat{\theta}^{(j)} \hat{f}_i(\hat{s}^{(j)})(1-\hat{s}_1^{(j)}) \right) \right. \\ &\quad \left. + \int_{\hat{T}} \hat{f}_4(\hat{y}) d\hat{y} - \sum_{j=1}^M \hat{q}^{(j)} \hat{f}_4(\hat{y}^{(j)}) \right). \end{aligned}$$

It then follows from the Bramble–Hilbert lemma that

$$\begin{aligned} E_T(f) &\leq C|T| \left( \sum_{i=1}^3 |(1 - \hat{s}_1) \hat{f}_i|_{H^{m+1}(\hat{Q})} + |\hat{f}_4|_{H^{m+1}(\hat{T})} \right) \\ &\leq Ch_T^2 \left( \sum_{i=1}^3 |\hat{f}_i|_{H^{m+1}(\hat{Q})} + |\hat{f}_4|_{H^{m+1}(\hat{T})} \right). \quad \square \end{aligned}$$

**Corollary 4.4.** *Let  $B^{(j)}$  be any of the three rational edge bubbles defined on an arbitrary triangle  $T \in \mathcal{T}_h$ . Then for any multi-index  $\alpha = (\alpha_1, \alpha_2)$  with  $0 \leq |\alpha| \leq 2$ , there exists  $C_\alpha > 0$  and  $\delta_\alpha > 0$  such that*

$$\left| E_T \left( \frac{\partial^{|\alpha|} B^{(j)}}{\partial \hat{x}^\alpha} \right) \right| \leq h_T^2 C_\alpha (\exp(-\delta_\alpha M) + \exp(-\delta_\alpha L)).$$

*Proof.* This follows directly from Theorem 4.2 and Theorem 4.3. □

### 5. Numerical experiments on a single triangle

In this section, we implement the quadrature scheme discussed in the previous section on the reference triangle  $\hat{T}$  and validate the results of Theorem 4.2. In all of the numerical experiments, we approximate the integral of the third function in (2-4), that is,

$$B(x) := B^{(3)}(x) = \frac{x_1^2 x_2^2 (1 - x_1 - x_2)}{(1 - x_1)(1 - x_2)}.$$

For comparison, we first implement some standard Gauss–Legendre quadrature schemes for the rational function and its first and second derivatives. Using the mathematical software package Maple, we find the exact value of the integrals to be

$$\begin{aligned} \int_{\hat{T}} B(\hat{x}) \, d\hat{x} &= -\frac{1}{6}\pi^2 + \frac{593}{360} \approx 0.0022881548227867501, \\ \int_{\hat{T}} \frac{\partial B}{\partial \hat{x}_1} \, d\hat{x} &= 0, \quad \int_{\hat{T}} \frac{\partial^2 B}{\partial \hat{x}_1^2} \, d\hat{x} = -\frac{1}{6}. \end{aligned}$$

The numerical results are depicted in Table 1. As can be seen from the tables, the errors behave sporadically. At best, the errors converge algebraically, but certainly not exponentially. Moreover, even for high order quadrature rules, we are only able to recover four digits of accuracy. This also proves true for Gaussian quadrature applied to the function’s first and second derivatives (see Table 1). We can attribute these poor results to the two singularities at the vertices  $(1, 0)$  and  $(0, 1)$ .

Next, we implement the quadrature scheme using the Duffy transform described in Section 4. Of particular interest are the integrals on the subtriangles containing

	$L$	degree	approx. integral	absolute error	relative error	rate
$\hat{B}$	1	1	$4.62 \cdot 10^{-3}$	$2.34 \cdot 10^{-3}$	1.023	
	3	2	$-9.74 \cdot 10^{-2}$	$9.96 \cdot 10^{-2}$	4.35	3.41
	6	4	$2.36 \cdot 10^{-3}$	$7.83 \cdot 10^{-5}$	$3.42 \cdot 10^{-2}$	10.31
	7	5	$2.29 \cdot 10^{-3}$	$3.51 \cdot 10^{-6}$	$1.53 \cdot 10^{-3}$	20.15
	16	8	$2.28 \cdot 10^{-3}$	$1.71 \cdot 10^{-6}$	$7.51 \cdot 10^{-4}$	0.86
	19	9	$2.28 \cdot 10^{-3}$	$2.81 \cdot 10^{-7}$	$1.22 \cdot 10^{-4}$	10.54
	28	11	$2.28 \cdot 10^{-3}$	$1.21 \cdot 10^{-7}$	$5.29 \cdot 10^{-5}$	2.17
	37	13	$2.30 \cdot 10^{-3}$	$2.08 \cdot 10^{-5}$	$9.11 \cdot 10^{-3}$	18.48
1st derivative of $\hat{B}$	1	1	$-2.08 \cdot 10^{-2}$	$2.08 \cdot 10^{-2}$		
	3	2	$8.78 \cdot 10^{-1}$	$8.78 \cdot 10^{-1}$		5.39
	6	4	$-2.02 \cdot 10^{-3}$	$2.02 \cdot 10^{-3}$		8.76
	7	5	$-1.27 \cdot 10^{-3}$	$1.27 \cdot 10^{-3}$		2.08
	16	8	$1.05 \cdot 10^{-4}$	$1.05 \cdot 10^{-4}$		5.29
	19	9	$1.02 \cdot 10^{-4}$	$1.02 \cdot 10^{-4}$		0.24
	28	11	$8.88 \cdot 10^{-6}$	$8.88 \cdot 10^{-6}$		12.18
	37	13	$4.12 \cdot 10^{-4}$	$4.12 \cdot 10^{-4}$		22.97
2nd derivative of $\hat{B}$	1	1	$5.80 \cdot 10^{-2}$	$2.18 \cdot 10^{-1}$	-1.31	
	3	2	-7.13	6.97	-41.83	4.99
	6	4	$-1.39 \cdot 10^{-1}$	$2.69 \cdot 10^{-2}$	$-1.61 \cdot 10^{-1}$	8.01
	7	5	$-1.45 \cdot 10^{-1}$	$2.09 \cdot 10^{-2}$	-1.25	1.12
	16	8	$-1.62 \cdot 10^{-1}$	$4.06 \cdot 10^{-3}$	$-2.43 \cdot 10^{-2}$	3.49
	19	9	$-1.62 \cdot 10^{-1}$	$3.70 \cdot 10^{-3}$	$-2.22 \cdot 10^{-2}$	0.78
	28	11	$-1.65 \cdot 10^{-1}$	$8.41 \cdot 10^{-4}$	$-5.05 \cdot 10^{-3}$	7.37
	37	13	$-1.75 \cdot 10^{-1}$	$8.60 \cdot 10^{-3}$	$-5.16 \cdot 10^{-2}$	13.91

**Table 1.** Gaussian quadrature results for the function  $\hat{B}$  of (3-8) and its first two derivatives. Rates of convergence are with respect to the relative error and the number of points  $L$ .

the singularities. For the sake of brevity we omit  $\hat{B}$  and its derivatives over the subtriangle  $\hat{K}_4 = \{(0, 5, 0), (0.5, 0.5), (0.5, 0)\}$ ; that is, we approximate the integrals

$$\int_{\hat{K}_1} B(\hat{x}) dx = -\frac{2}{3} \ln 2 + \frac{6019}{5760} - \frac{1}{12} \pi^2 + \frac{1}{2} \ln^2 2 \approx 6.266309395 \times 10^{-4},$$

$$\int_{\hat{K}_1} \frac{\partial B}{\partial \hat{x}_1} dx = -\frac{17}{96} + \frac{1}{4} \ln 2 \approx -3.7955381933 \times 10^{-3},$$

$$\int_{\hat{K}_1} \frac{\partial^2 B}{\partial \hat{x}_1^2} dx = -\frac{35}{24} + \ln 4 \approx -7.20389722 \times 10^{-2},$$

	$L$	degree	approx. integral	absolute error	relative error	rate
$\hat{B}$	4	2	$6.66 \cdot 10^{-4}$	$3.90 \cdot 10^{-5}$	$6.22 \cdot 10^{-2}$	
	9	3	$6.27 \cdot 10^{-4}$	$4.00 \cdot 10^{-7}$	$6.38 \cdot 10^{-4}$	5.65
	16	4	$6.27 \cdot 10^{-4}$	$9.40 \cdot 10^{-9}$	$1.50 \cdot 10^{-5}$	6.52
	25	5	$6.27 \cdot 10^{-4}$	$2.23 \cdot 10^{-10}$	$3.57 \cdot 10^{-7}$	8.38
	36	6	$6.27 \cdot 10^{-4}$	$5.48 \cdot 10^{-12}$	$8.74 \cdot 10^{-9}$	10.17
	49	7	$6.27 \cdot 10^{-4}$	$1.38 \cdot 10^{-13}$	$2.20 \cdot 10^{-10}$	11.95
	64	8	$6.27 \cdot 10^{-4}$	$9.10 \cdot 10^{-17}$	$1.45 \cdot 10^{-13}$	27.42
1st derivative of $\hat{B}$	4	2	$-3.58 \cdot 10^{-3}$	$2.12 \cdot 10^{-4}$	$-5.60 \cdot 10^{-2}$	
	9	3	$-3.79 \cdot 10^{-3}$	$3.76 \cdot 10^{-6}$	$-9.92 \cdot 10^{-4}$	4.97
	16	4	$-3.79 \cdot 10^{-3}$	$7.95 \cdot 10^{-8}$	$-2.09 \cdot 10^{-5}$	6.70
	25	5	$-3.79 \cdot 10^{-3}$	$1.82 \cdot 10^{-9}$	$-4.81 \cdot 10^{-7}$	8.45
	36	6	$-3.79 \cdot 10^{-3}$	$4.41 \cdot 10^{-11}$	$-1.16 \cdot 10^{-8}$	10.21
	49	7	$-3.79 \cdot 10^{-3}$	$1.10 \cdot 10^{-12}$	$-2.90 \cdot 10^{-10}$	11.97
	64	8	$-3.79 \cdot 10^{-3}$	$2.81 \cdot 10^{-14}$	$-7.40 \cdot 10^{-12}$	13.73
2nd derivative of $\hat{B}$	4	2	$-7.07 \cdot 10^{-2}$	$1.24 \cdot 10^{-3}$	$-1.72 \cdot 10^{-2}$	
	9	3	$-7.20 \cdot 10^{-2}$	$2.34 \cdot 10^{-5}$	$-3.24 \cdot 10^{-4}$	4.89
	16	4	$-7.20 \cdot 10^{-2}$	$5.10 \cdot 10^{-7}$	$-7.08 \cdot 10^{-6}$	6.64
	25	5	$-7.20 \cdot 10^{-2}$	$2.91 \cdot 10^{-10}$	$-1.65 \cdot 10^{-7}$	8.41
	36	6	$-7.20 \cdot 10^{-2}$	$7.34 \cdot 10^{-12}$	$-4.04 \cdot 10^{-9}$	10.18
	49	7	$-7.20 \cdot 10^{-2}$	$1.88 \cdot 10^{-13}$	$-1.01 \cdot 10^{-10}$	11.94
	64	8	$-7.20 \cdot 10^{-2}$	$4.99 \cdot 10^{-15}$	$-2.62 \cdot 10^{-12}$	13.70

**Table 2.** Quadrature results using the Duffy transform for the function  $\hat{B}$  of (3-8) and its first two derivatives. The domain of integration is the triangle  $\hat{K}_1$  (the next two tables deal with  $\hat{K}_2$  and  $\hat{K}_3$ ). Rates of convergence are with respect to the relative error and the number of points  $L$ .

$$\int_{\hat{K}_2} B(\hat{x}) dx = -\frac{2}{3} \ln 2 + \frac{6019}{5760} - \frac{1}{12} \pi^2 + \frac{1}{2} \ln^2 2 \approx 6.266309395 \times 10^{-4},$$

$$\int_{\hat{K}_2} \frac{\partial^2 B}{\partial \hat{x}_1} dx = 0,$$

$$\int_{\hat{K}_2} \frac{\partial^2 B}{\partial \hat{x}_1^2} dx = -\frac{1}{12} \approx -0.08333333333,$$

$$\int_{\hat{K}_3} B(\hat{x}) dx \approx 8.096731144 \times 10^{-5},$$

$$\int_{\hat{K}_3} \frac{\partial B}{\partial \hat{x}_1} dx = \frac{1}{6} \ln 2 - \frac{11}{96} \approx 9.411967600 \times 10^{-4},$$

	$L$	degree	approx. integral	absolute error	relative error	rate
$\hat{B}$	4	2	$6.66 \cdot 10^{-4}$	$3.89 \cdot 10^{-5}$	$6.22 \cdot 10^{-2}$	
	9	3	$6.27 \cdot 10^{-4}$	$3.99 \cdot 10^{-7}$	$6.38 \cdot 10^{-4}$	5.64
	16	4	$6.26 \cdot 10^{-4}$	$9.40 \cdot 10^{-9}$	$1.50 \cdot 10^{-5}$	6.51
	25	5	$6.26 \cdot 10^{-4}$	$2.23 \cdot 10^{-11}$	$3.57 \cdot 10^{-7}$	8.36
	36	6	$6.26 \cdot 10^{-4}$	$5.47 \cdot 10^{-12}$	$8.74 \cdot 10^{-9}$	10.17
	49	7	$6.26 \cdot 10^{-4}$	$1.37 \cdot 10^{-13}$	$2.20 \cdot 10^{-10}$	11.94
	64	8	$6.26 \cdot 10^{-4}$	$3.52 \cdot 10^{-15}$	$1.45 \cdot 10^{-13}$	13.71
1st derivative of $\hat{B}$	4	2	$8.61 \cdot 10^{-4}$	$8.61 \cdot 10^{-4}$		
	9	3	$4.25 \cdot 10^{-6}$	$4.25 \cdot 10^{-6}$		6.54
	16	4	$1.19 \cdot 10^{-7}$	$1.19 \cdot 10^{-7}$		6.20
	25	5	$3.43 \cdot 10^{-9}$	$3.43 \cdot 10^{-9}$		7.96
	36	6	$9.89 \cdot 10^{-11}$	$9.89 \cdot 10^{-11}$		9.72
	49	7	$2.86 \cdot 10^{-13}$	$2.86 \cdot 10^{-3}$		11.48
	64	8	$8.33 \cdot 10^{-15}$	$8.33 \cdot 10^{-15}$		13.24
2nd derivative of $\hat{B}$	4	2	$-8.21 \cdot 10^{-2}$	$1.00 \cdot 10^{-3}$	$-1.44 \cdot 10^{-2}$	
	9	3	$-8.32 \cdot 10^{-2}$	$4.55 \cdot 10^{-5}$	$-5.46 \cdot 10^{-4}$	4.03
	16	4	$-8.33 \cdot 10^{-2}$	$1.63 \cdot 10^{-6}$	$-1.95 \cdot 10^{-5}$	5.78
	25	5	$-8.33 \cdot 10^{-2}$	$5.64 \cdot 10^{-8}$	$-6.77 \cdot 10^{-7}$	7.53
	36	6	$-8.33 \cdot 10^{-2}$	$1.90 \cdot 10^{-9}$	$-2.29 \cdot 10^{-8}$	9.28
	49	7	$-8.33 \cdot 10^{-2}$	$6.34 \cdot 10^{-11}$	$-7.61 \cdot 10^{-10}$	11.04
	64	8	$-8.33 \cdot 10^{-2}$	$2.08 \cdot 10^{-13}$	$-2.49 \cdot 10^{-11}$	12.79

**Table 3.** Quadrature results over the domain of integration  $\hat{K}_2$ . See caption of Table 2 for details.

$$\int_{\hat{K}_3} \frac{\partial^2 B}{\partial \hat{x}_1^2} dx = \frac{5}{8} - \frac{8}{9} \ln 2 \approx 8.869172836 \times 10^{-3}$$

by the quadrature scheme  $\sum_{j=1}^L \hat{\omega}^{(j)} B(\hat{x}^{(j)})$ . The numerical results and the errors are listed in the tables below. Our error now converges in an exponential manner for our initial function as well as its first and second derivative. These results are in agreement with Theorem 4.2.

### 6. Conclusion

In this paper, we have created an effective Gaussian quadrature scheme for a specific class of divergence free rational functions. We also managed to derive error estimates as well as show exponential error convergence, with numerical experiments confirming our results. While the findings of this paper appear to support the finite element method proposed in [Guzmán and Neilan 2014a], there

	$L$	degree	approx. integral	absolute error	relative error	rate
$\hat{B}$	4	2	$6.35 \cdot 10^{-5}$	$1.73 \cdot 10^{-5}$	$2.14 \cdot 10^{-1}$	
	9	3	$8.08 \cdot 10^{-5}$	$1.35 \cdot 10^{-7}$	$1.67 \cdot 10^{-3}$	5.98
	16	4	$8.09 \cdot 10^{-5}$	$3.60 \cdot 10^{-9}$	$4.45 \cdot 10^{-5}$	6.30
	25	5	$8.09 \cdot 10^{-5}$	$1.12 \cdot 10^{-10}$	$1.38 \cdot 10^{-6}$	7.77
	36	6	$8.09 \cdot 10^{-5}$	$3.06 \cdot 10^{-12}$	$3.78 \cdot 10^{-8}$	9.88
	49	7	$8.09 \cdot 10^{-5}$	$4.43 \cdot 10^{-13}$	$5.47 \cdot 10^{-10}$	13.73
	64	8	$8.09 \cdot 10^{-5}$	$3.56 \cdot 10^{-13}$	$4.39 \cdot 10^{-10}$	.82
1st derivative of $\hat{B}$	4	2	$9.47 \cdot 10^{-4}$	$6.04 \cdot 10^{-6}$	$6.42 \cdot 10^{-3}$	
	9	3	$9.42 \cdot 10^{-4}$	$1.04 \cdot 10^{-6}$	$1.10 \cdot 10^{-3}$	4.16
	16	4	$9.41 \cdot 10^{-4}$	$1.00 \cdot 10^{-8}$	$1.07 \cdot 10^{-5}$	8.06
	25	5	$9.41 \cdot 10^{-4}$	$1.09 \cdot 10^{-10}$	$1.16 \cdot 10^{-7}$	10.12
	36	6	$9.41 \cdot 10^{-4}$	$9.74 \cdot 10^{-12}$	$1.03 \cdot 10^{-8}$	6.64
	49	7	$9.41 \cdot 10^{-4}$	$3.97 \cdot 10^{-13}$	$4.22 \cdot 10^{-10}$	10.37
	64	8	$9.41 \cdot 10^{-4}$	$1.38 \cdot 10^{-14}$	$1.46 \cdot 10^{-11}$	12.58
2nd derivative of $\hat{B}$	4	2	$9.00 \cdot 10^{-3}$	$1.34 \cdot 10^{-4}$	$1.51 \cdot 10^{-2}$	
	9	3	$8.87 \cdot 10^{-3}$	$2.95 \cdot 10^{-6}$	$3.33 \cdot 10^{-4}$	4.70
	16	4	$8.86 \cdot 10^{-3}$	$6.96 \cdot 10^{-8}$	$7.85 \cdot 10^{-6}$	6.51
	25	5	$8.86 \cdot 10^{-3}$	$6.21 \cdot 10^{-9}$	$7.00 \cdot 10^{-7}$	5.41
	36	6	$8.86 \cdot 10^{-3}$	$2.85 \cdot 10^{-11}$	$3.22 \cdot 10^{-8}$	8.44
	49	7	$8.86 \cdot 10^{-3}$	$1.11 \cdot 10^{-12}$	$1.25 \cdot 10^{-9}$	10.53
	64	8	$8.86 \cdot 10^{-3}$	$4.01 \cdot 10^{-14}$	$4.53 \cdot 10^{-11}$	12.42

**Table 4.** Quadrature results over the domain of integration  $\hat{K}_2$ . See caption of Table 2 for details.

are still a number of conditions such as  $V_h$  [Ciarlet 1978, p. 174], ellipticity and determining global error estimates which must be worked out. This will be the subject of ongoing research.

### Acknowledgements

The author would like to thank his advisor, Dr. Michael Neilan, for suggesting this problem and for thoughtful discussions.

### References

- [Brezzi and Fortin 1991] F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*, Springer Series in Computational Mathematics **15**, Springer, New York, 1991. MR 92d:65187 Zbl 0788.73002
- [Burden and Faries 2011] R. Burden and J. Faries, *Numerical analysis*, 9th ed., Brooks/Cole, Cengage Learning, Pacific Grove, CA, 2011.



- [Ciarlet 1978] P. G. Ciarlet, *The finite element method for elliptic problems*, Studies in Mathematics and its Applications **4**, North-Holland Publishing Co., Amsterdam, 1978. MR 58 #25001
- [Duffy 1982] M. G. Duffy, “Quadrature over a pyramid or cube of integrands with a singularity at a vertex”, *SIAM J. Numer. Anal.* **19**:6 (1982), 1260–1262. MR 83k:65020
- [Guzmán and Neilan 2014a] J. Guzmán and M. Neilan, “Conforming and divergence-free Stokes elements on general triangular meshes”, *Math. Comp.* **83**:285 (2014), 15–36. MR 3120580 Zbl 06227546
- [Guzmán and Neilan 2014b] J. Guzmán and M. Neilan, “Symmetric and conforming mixed finite elements for plane elasticity using rational bubble functions”, *Numer. Math.* **126**:1 (2014), 153–171. MR 3149075 Zbl 06261585
- [Lyness and Cools 1994] J. N. Lyness and R. Cools, “A survey of numerical cubature over triangles”, pp. 127–150 in *Mathematics of Computation 1943–1993: a half-century of computational mathematics* (Vancouver, BC, 1993), edited by W. Gautschi, Proc. Sympos. Appl. Math. **48**, Amer. Math. Soc., Providence, RI, 1994. MR 95j:65021
- [Sauter and Schwab 2011] S. A. Sauter and C. Schwab, *Boundary element methods*, Springer Series in Computational Mathematics **39**, Springer, Berlin, 2011. MR 2011i:65003
- [Spivak 1998] M. Spivak, *Calculus on Manifolds*, Perseus, Cambridge, MA, 1998. Reprint of the 1965 original.

Received: 2012-11-24    Revised: 2013-07-17    Accepted: 2013-07-29

mschneier89@gmail.com



# Finite groups with some weakly $s$ -permutably embedded and weakly $s$ -supplemented subgroups

Guo Zhong, XuanLong Ma, Shixun Lin, Jiayi Xia and Jianxing Jin

(Communicated by Joseph A. Gallian)

Let  $G$  be a finite group. A subgroup  $H$  of  $G$  is called weakly  $s$ -permutably embedded in  $G$  if there is a subnormal subgroup  $T$  of  $G$  and an  $s$ -permutably embedded subgroup  $H_{se}$  of  $G$  contained in  $H$  such that  $G = HT$  and  $H \cap T \leq H_{se}$ . The subgroup  $H$  is called weakly  $s$ -supplemented in  $G$  if  $G$  has a subgroup  $K$  such that  $HK = G$  and  $H \cap K \leq H_{sG}$ , where  $H_{sG}$  is the largest  $s$ -permutable subgroup of  $G$  contained in  $H$ . In this paper, we investigate the influence of weakly  $s$ -permutably embedded and weakly  $s$ -supplemented subgroups on the structure of finite groups. Some recent results are generalized.

## 1. Introduction

Throughout only finite groups are considered. We use conventional terminology and notation, as in [Robinson 1982]. Let  $G$  denote a group and  $|G|$  denote the order of  $G$ . Let  $B \trianglelefteq A \leq G$ . Then  $A/B$  is a section of  $G$ . In the theory of groups,  $G$  is said to be  $A_4$ -free if  $G$  does not possess a section isomorphic to  $A_4$ .

Let  $\mathcal{F}$  be a class of groups. Then  $\mathcal{F}$  is called a formation provided that (1) if  $G \in \mathcal{F}$  and  $H \triangleleft G$ , then  $G/H \in \mathcal{F}$ , and (2) if  $G/M$  and  $G/N$  are in  $\mathcal{F}$ , then  $G/M \cap N$  is in  $\mathcal{F}$  for all normal subgroups  $M, N$  of  $G$ . A formation  $\mathcal{F}$  is said to be saturated if  $G/\Phi(G) \in \mathcal{F}$  implies that  $G \in \mathcal{F}$ , where  $\Phi(G)$  denotes the Frattini subgroup of  $G$ .

Two subgroups  $H$  and  $K$  of  $G$  are said to be permutable if  $HK = KH$ . Following [Kegel 1962], the subgroup  $H$  of  $G$  is said to be  $s$ -permutable in  $G$  if  $H$  permutes with every Sylow subgroup of  $G$ , that is,  $HP = PH$  for any Sylow subgroup  $P$  of  $G$ . Schmid [1998] showed that if both  $H$  and  $K$  are  $s$ -permutable subgroups of  $G$ , then both  $H \cap K$  and  $\langle H, K \rangle$  are  $s$ -permutable in  $G$ . Recently, Ballester-Bolinches and Pedraza-Aguilera [1998] generalized  $s$ -permutable subgroups to  $s$ -permutably

*MSC2010:* primary 20D10; secondary 20D20.

*Keywords:* weakly  $s$ -permutably embedded subgroups, weakly  $s$ -supplemented subgroups,  $p$ -nilpotent groups.

embedded subgroups. A subgroup  $H$  is said to be  $s$ -permutably embedded in  $G$  provided every Sylow subgroup of  $H$  is a Sylow subgroup of some  $s$ -permutable subgroup of  $G$ . By applying this concept, Ballester-Bolinches and Pedraza-Aguilera got new criteria for the supersolvability of groups. Moreover, a nice result in [Li et al. 2005] on the  $p$ -nilpotency of a group could be stated as follows: Let  $G$  be a group and  $P$  a Sylow  $p$ -subgroup of  $G$ , where  $p$  is the smallest prime dividing  $|G|$ . If  $G$  is  $A_4$ -free and all 2-maximal subgroups of  $P$  are  $s$ -permutably embedded in  $G$ , then  $G$  is  $p$ -nilpotent.

In recent years, it has been of interest to use supplementation properties of subgroups to characterize properties of a group. Wang [1996] first introduced the concept of  $c$ -normal subgroups. Furthermore, Li, Qiao, and Wang [Li et al. 2009] continued to promote this concept and introduced weakly  $s$ -permutably embedded subgroups, which are a generalization of both  $c$ -normality [Wang 1996] and  $s$ -permutably embedding. A subgroup  $H$  of  $G$  is called weakly  $s$ -permutably embedded in  $G$  if there is a subnormal subgroup  $T$  of  $G$  and an  $s$ -permutably embedded subgroup  $H_{se}$  of  $G$  contained in  $H$  such that  $G = HT$  and  $H \cap T \leq H_{se}$ . In the meantime, Skiba [2007] introduced the definition of a weakly  $s$ -supplemented subgroup. A subgroup  $H$  is said to be weakly  $s$ -supplemented in  $G$  if  $G$  has a subgroup  $T$  such that  $HT = G$  and  $H \cap T \leq H_{sG}$ , where  $H_{sG}$  is the largest  $s$ -permutable subgroup of  $G$  contained in  $H$ .

We note that weakly  $s$ -permutably embedded subgroups and weakly  $s$ -supplemented subgroups are two distinct concepts. There are examples that show that weakly  $s$ -permutably embedded subgroups are not weakly  $s$ -supplemented subgroups, and, in general, the converse is also false.

**Example 1.1.** Let  $G = A_5$  be the alternating group of degree 5. Then the Sylow 2-subgroups of  $G$  are weakly  $s$ -permutably embedded in  $G$ , but not weakly  $s$ -supplemented in  $G$ .

**Example 1.2.** Let  $H = S_4$  be the symmetric group of degree 4, let  $V$  be an irreducible and faithful module for  $H$  over  $\mathbb{F}_3$ , the finite field of 3 elements, and consider  $G = [V]H$ , the corresponding semidirect product. If  $X$  is a Sylow 3-subgroup of  $H$ , then  $X$  is weakly  $s$ -supplemented in  $G$  but not weakly  $s$ -permutably embedded in  $G$ .

Hence it is natural to ask the following question: can these two concepts and the related results be unified and generalized? The purpose of this article is to present an answer to the above question. By using these subgroup properties, we determine the structure of  $G$  based on the assumption that all 2-maximal subgroups of a Sylow subgroup of  $G$  are either weakly  $s$ -permutably embedded or weakly  $s$ -supplemented subgroups in  $G$ . Our results unify and generalize the above mentioned result and some other results in the literature on  $p$ -nilpotency and formation theory of finite groups.

## 2. Preliminaries

For the sake of convenience, we include the following results.

**Lemma 2.1** [Ballester-Bolinches and Pedraza-Aguilera 1998, Lemma 1]. *Let  $H$  be a subgroup of  $G$ .*

- (1) *If  $H$  is  $s$ -permutably embedded in  $G$  and  $H \leq M \leq G$ , then  $H$  is  $s$ -permutably embedded in  $M$ .*
- (2) *Let  $N \triangleleft G$  and assume that  $H$  is  $s$ -permutably embedded in  $G$ . Then  $HN$  is  $s$ -permutably embedded in  $G$  and  $HN/N$  is  $s$ -permutably embedded in  $G/N$ .*

**Lemma 2.2** [Li et al. 2009, Lemma 2.5]. *Let  $U$  be a weakly  $s$ -permutably embedded subgroup of  $G$  and  $N$  a normal subgroup of  $G$ . Then:*

- (1) *If  $U \leq H \leq G$ , then  $U$  is weakly  $s$ -permutably embedded in  $H$ .*
- (2) *If  $N \leq U$ , then  $U/N$  is weakly  $s$ -permutably embedded in  $G/N$ .*
- (3) *Let  $\pi$  be a set of primes,  $U$  a  $\pi$ -subgroup and  $N$  a  $\pi'$ -subgroup. Then  $UN/N$  is weakly  $s$ -permutably embedded in  $G/N$ .*

**Lemma 2.3** [Skiba 2007, Lemma 2.10]. *Let  $H$  be a subgroup of a group  $G$ .*

- (1) *If  $H$  is weakly  $s$ -supplemented in  $G$  and  $H \leq M \leq G$ , then  $H$  is weakly  $s$ -supplemented in  $M$ .*
- (2) *Let  $N \triangleleft G$  and  $N \leq H$ . If  $H$  is weakly  $s$ -supplemented in  $G$ , then  $H/N$  is weakly  $s$ -supplemented in  $G/N$ .*
- (3) *Let  $\pi$  be a set of primes,  $H$  a  $\pi$ -subgroup of  $G$  and  $N$  a normal  $\pi'$ -subgroup of  $G$ . If  $H$  is weakly  $s$ -supplemented in  $G$ , then  $HN/N$  is weakly  $s$ -supplemented in  $G/N$ .*

**Lemma 2.4** [Guo and Shum 2003, Lemma 3.12]. *Let  $P$  be a Sylow  $p$ -subgroup of a group  $G$ , where  $p$  is the smallest prime dividing  $|G|$ . If  $G$  is  $A_4$ -free and  $|P| \leq p^2$ , then  $G$  is  $p$ -nilpotent.*

**Lemma 2.5** [Guo et al. 2009, Lemma 2.12]. *Let  $p$  be a prime, and let  $G$  be a group with  $(|G|, p-1) = 1$ . Suppose that  $P$  is a Sylow  $p$ -subgroup of  $G$  such that each maximal subgroup of  $P$  has a  $p$ -nilpotent supplement in  $G$ . Then  $G$  is  $p$ -nilpotent.*

**Lemma 2.6** [Li et al. 2005]. (1) *If  $P$  is an  $s$ -permutable  $p$ -subgroup of  $G$  for some prime  $p$ , then  $O^p(G) \leq N_G(P)$ .*

- (2) *Suppose that  $H$  is  $s$ -permutable in  $G$  and  $P$  is a Sylow  $p$ -subgroup of  $H$ , where  $p$  is a prime. If  $H_G = 1$ , then  $P$  is  $s$ -permutable in  $G$ .*
- (3) *Suppose that  $P$  is a  $p$ -subgroup of  $G$  contained in  $O_p(G)$ . If  $P$  is  $s$ -permutably embedded in  $G$ , then  $P$  is  $s$ -permutable in  $G$ .*

**Lemma 2.7** [Li and Guo 2000, Lemma 2.6]. *Let  $H$  be a nontrivial solvable normal subgroup of  $G$ . If every minimal normal subgroup of  $G$  which is contained in  $H$  is not contained in  $\Phi(G)$ , then the Fitting subgroup  $F(H)$  of  $H$  is the direct product of minimal normal subgroups of  $G$  which are contained in  $H$ .*

**Lemma 2.8** [Doerk and Hawkes 1992, A, Lemma 1.2]. *Let  $U, V$  and  $W$  be subgroups of  $G$ . The following statements are equivalent:*

- (1)  $U \cap VW = (U \cap V)(U \cap W)$ .
- (2)  $UV \cap UW = U(V \cap W)$ .

**Lemma 2.9** [Guo and Shum 2003, Lemma 3.16]. *Let  $\mathcal{F}$  be the class of groups with Sylow tower of supersolvable type. Also let  $P$  be a normal  $p$ -subgroup of  $G$  such that  $G/P \in \mathcal{F}$ . If  $G$  is  $A_4$ -free and  $|P| \leq p^2$ , then  $G \in \mathcal{F}$ .*

**Lemma 2.10** [Zhang and Li 2012, Lemma 2.11]. *Let  $p$  be the smallest prime dividing  $|G|$  and  $P$  a Sylow  $p$ -subgroup of  $G$ . If  $G$  is  $A_4$ -free and every 2-maximal subgroup of  $P$  is weakly  $s$ -permutably embedded in  $G$ , then  $G$  is  $p$ -nilpotent.*

**Lemma 2.11** [Yang et al. 2012, Lemma 2.12]. *If a  $p$ -subgroup  $H$  is  $s$ -permutable in  $G$ , then  $H \leq O_p(G)$ .*

### 3. Main results

Our first result unifies and improves the results [Ballester-Bolinchés and Guo 1999, Theorem 3; Guo and Shum 2001, Theorem 3.2; Wang 2000, Theorem 4.2] on the  $p$ -nilpotency of a group.

**Theorem 3.1.** *Let  $p$  be the smallest prime dividing  $|G|$  and  $P$  a Sylow  $p$ -subgroup of  $G$ . If  $G$  is  $A_4$ -free and every 2-maximal subgroup of  $P$  is either weakly  $s$ -permutably embedded or weakly  $s$ -supplemented in  $G$ , then  $G$  is  $p$ -nilpotent.*

*Proof.* Suppose that the statement is false and let  $G$  be a counterexample of minimal order. We proceed with the following steps.

Step 1: By Lemma 2.4,  $|P| \geq p^3$  and thus every 2-maximal subgroup of  $P$  is nontrivial.

Step 2:  $G$  is not a nonabelian simple group.

Assume that  $G$  is nonabelian simple. By Lemma 2.5,  $P$  has a maximal subgroup  $P_1$  which has no  $p$ -nilpotent supplement in  $G$ . It follows that any 2-maximal subgroup  $P_2$  of  $P$  contained in  $P_1$  has no  $p$ -nilpotent supplement in  $G$ . From the hypothesis,  $P_2$  is either weakly  $s$ -permutably embedded or weakly  $s$ -supplemented in  $G$ . If  $P_2$  is weakly  $s$ -permutably embedded in  $G$ , then there is a subnormal subgroup  $T$  of  $G$  and an  $s$ -permutably embedded subgroup  $(P_2)_{se}$  of  $G$  contained in  $P_2$  such that  $G = P_2T$  and  $P_2 \cap T \leq (P_2)_{se}$ . Clearly,  $T = G$  and thus  $P_2 = (P_2)_{se}$

is  $s$ -permutably embedded in  $G$ . Thus there is an  $s$ -permutable subgroup  $K$  of  $G$  such that  $P_2$  is a Sylow  $p$ -subgroup of  $K$ . Since  $G$  is simple, we get  $K_G = 1$ . By Lemma 2.6,  $P_2$  is  $s$ -permutable in  $G$ . Consequently,  $1 \neq P_2 \leq O_p(G)$  by Lemma 2.11, which is a contradiction. If  $P_2$  is weakly  $s$ -supplemented in  $G$ , then there is a non- $p$ -nilpotent subgroup  $T$  of  $G$  such that

$$G = P_2T \quad \text{and} \quad P_2 \cap T \leq (P_2)_{sG} \leq O_p(G) = 1$$

by Lemma 2.11. By Lemma 2.4,  $T$  is  $p$ -nilpotent, a contradiction.

Step 3:  $G$  has a unique minimal normal subgroup  $N$ , and  $G/N$  is  $p$ -nilpotent. Furthermore,  $\Phi(G) = 1$ .

Let  $N$  be a minimal normal subgroup of  $G$ . Consider the factor group  $G/N$ ; we will prove that  $G/N$  meets the hypotheses of the theorem. Since  $P$  is a Sylow  $p$ -subgroup of  $G$ ,  $PN/N$  is a Sylow  $p$ -subgroup of  $G/N$ . If  $|PN/N| \leq p^2$ , then  $G/N$  is  $p$ -nilpotent by Lemma 2.4. Hence we assume  $|PN/N| \geq p^3$ . Let  $M_2/N$  be a 2-maximal subgroup of  $PN/N$ . Then  $M_2 = N(M_2 \cap P)$ . Let  $P_2 = M_2 \cap P$ . It follows that  $P_2 \cap N = M_2 \cap P \cap N = P \cap N$  is a Sylow  $p$ -subgroup of  $N$ . Since

$$p^2 = |PN/N : M_2/N| = |PN : (M_2 \cap P)N| = |P : M_2 \cap P| = |P : P_2|,$$

$P_2$  is a 2-maximal subgroup of  $P$ . If  $P_2$  is weakly  $s$ -supplemented in  $G$ , then there is a subgroup  $T$  of  $G$  such that  $G = P_2T$  and  $P_2 \cap T \leq (P_2)_{sG}$ . So

$$G/N = M_2/N \cdot TN/N = P_2N/N \cdot TN/N.$$

Since  $(|N : P_2 \cap N|, |N : T \cap N|) = 1$ ,

$$(P_2 \cap N)(T \cap N) = N = N \cap G = N \cap P_2T.$$

By Lemma 2.8,  $(P_2N) \cap (TN) = (P_2 \cap T)N$ . It follows that

$$(P_2N/N) \cap (TN/N) = (P_2N \cap TN)/N = (P_2 \cap T)N/N \leq (P_2)_{sG}N/N.$$

By Lemma 2.6(2) of [Skiba 2007], we know that  $(P_2)_{sG}N/N$  is  $s$ -permutable in  $G$  and thus  $(P_2)_{sG}N/N \leq (P_2N/N)_{sG}$ . Hence  $M_2/N$  is weakly  $s$ -supplemented in  $G/N$ . If  $P_2$  is weakly  $s$ -permutably embedded in  $G$ , by Lemma 2.1, it follows analogously that  $M_2/N$  is weakly  $s$ -permutably embedded in  $G/N$ , too. Consequently,  $G/N$  meets the hypotheses of the theorem. The minimal choice of  $G$  implies that  $G/N$  is  $p$ -nilpotent. The uniqueness of  $N$  and  $\Phi(G) = 1$  are clear.

Step 4:  $O_{p'}(G) = 1$ .

If  $O_{p'}(G) \neq 1$ , then  $N \leq O_{p'}(G)$  by Step 3. Since

$$G/O_{p'}(G) \cong (G/N)/(O_{p'}(G)/N)$$

is  $p$ -nilpotent, we get that  $G$  is  $p$ -nilpotent, a contradiction.

Step 5:  $O_p(G) = 1$ .

If  $O_p(G) \neq 1$ , Step 3 yields  $N \leq O_p(G)$  and  $\Phi(O_p(G)) \leq \Phi(G) = 1$ . Hence,  $G$  has a maximal subgroup  $M$  such that  $G = MN$  and  $G/N \cong M$  is  $p$ -nilpotent. Since  $O_p(G) \cap M$  is normalized by  $N$  and  $M$ , and also by  $G$ , the uniqueness of  $N$  yields  $N = O_p(G)$ . Obviously,  $P = N(P \cap M)$ . Since  $P \cap M < P$ , there exists a maximal subgroup  $P_1$  of  $P$  such that  $P \cap M \leq P_1$ . Then  $P = NP_1$ . Pick a 2-maximal subgroup  $P_2$  of  $P$  such that  $P_2 \leq P_1$ . Under the hypothesis,  $P_2$  is either weakly  $s$ -permutably embedded or weakly  $s$ -supplemented in  $G$ . If  $P_2$  is weakly  $s$ -permutably embedded in  $G$ , then there is a subnormal subgroup  $T$  of  $G$  and an  $s$ -permutably embedded subgroup  $(P_2)_{se}$  of  $G$  contained in  $P_2$  such that  $G = P_2T$  and  $P_2 \cap T \leq (P_2)_{se}$ . Thus there is an  $s$ -permutable subgroup  $K$  of  $G$  such that  $(P_2)_{se}$  is a Sylow  $p$ -subgroup of  $K$ . If  $K_G \neq 1$ , then  $N \leq K_G \leq K$ . It follows that  $N \leq (P_2)_{se} \leq P_1$ , and thus  $P = N(P \cap M) = NP_1 = P_1$ , a contradiction. If  $K_G = 1$ , by Lemma 2.6,  $(P_2)_{se}$  is  $s$ -permutable in  $G$ . It follows from Lemma 2.11 that

$$P_2 \cap T \leq (P_2)_{se} \leq O_p(G) = N.$$

Hence,  $(P_2)_{se} \leq P_1 \cap N$ . It follows that

$$((P_2)_{se})^G = 1 \quad \text{or} \quad ((P_2)_{se})^G = P_1 \cap N = N.$$

If  $((P_2)_{se})^G = 1$ , then  $P_2 \cap T = 1$  and thus  $|T|_p = p^2$ . Hence  $T$  is  $p$ -nilpotent by Lemma 2.4. Let  $T_{p'}$  be the normal  $p$ -complement of  $T$ . Then  $T_{p'}$  is a normal Hall  $p'$ -subgroup of  $G$  since  $T$  is subnormal in  $G$ , which is a contradiction. If  $((P_2)_{se})^G = P_1 \cap N = N$ , then  $N \leq P_1$  and thus  $P = P_1$ , a contradiction. Now we may assume that  $P_2$  is weakly  $s$ -supplemented in  $G$ . Then there is a subgroup  $T$  of  $G$  such that  $G = P_2T$  and  $P_2 \cap T \leq (P_2)_{sG} \leq O_p(G) = N$  by Lemma 2.11. Similarly, we get that

$$((P_2)_{sG})^G = 1 \quad \text{or} \quad ((P_2)_{sG})^G = P_1 \cap N = N.$$

Arguing as before we may assume that  $((P_2)_{sG})^G = 1$  and deduce that  $T$  is  $p$ -nilpotent. Let  $T_{p'}$  be the normal  $p$ -complement of  $T$ . Since  $M$  is  $p$ -nilpotent, we have that  $M$  has a normal Hall  $p'$ -subgroup  $M_{p'}$  and  $M \leq N_G(M_{p'}) \leq G$ . The maximality of  $M$  and the fact that  $O_{p'}(G) = 1$  imply that  $M = N_G(M_{p'})$ . By using a deep result of Gross [1987, main theorem], there exists  $g \in G$  such that  $T_{p'}^g = M_{p'}$ . Hence  $T^g \leq N_G(T_{p'}^g) = N_G(M_{p'}) = M$ . But  $T_{p'}$  is normalized by  $T$ , thus  $g$  can be considered to be an element of  $P_2$ . It follows that  $G = P_2T^g = P_2M$  and  $P = P_2(P \cap M) = P_1$ , a contradiction.

Step 6:  $G$  has Hall  $p'$ -subgroups and any two Hall  $p'$ -subgroups of  $G$  are conjugate in  $G$ .



If every 2-maximal subgroup of  $P$  is weakly  $s$ -permutably embedded in  $G$ , then  $G$  is  $p$ -nilpotent by Lemma 2.10, a contradiction. Thus there is a 2-maximal subgroup  $P_2$  of  $P$  such that  $P_2$  is weakly  $s$ -supplemented in  $G$ . Then there exists a subgroup  $T$  of  $G$  such that  $G = P_2T$  and  $P_2 \cap T \leq (P_2)_{sG} \leq O_p(G) = 1$  by Lemma 2.11. By Lemma 2.4,  $T$  is  $p$ -nilpotent and thus  $T$  has normal  $p$ -complement  $T_{p'}$ . Obviously,  $T_{p'}$  is also a Hall  $p'$ -subgroup of  $G$ . By [Gross 1987, main theorem], we have that any two Hall  $p'$ -subgroups of  $G$  are conjugate in  $G$ .

Step 7: The final contradiction.

If  $NP < G$ , then  $NP$  meets the hypotheses of the theorem. The minimal choice of  $G$  yields that  $NP$  is  $p$ -nilpotent. Let  $N_{p'}$  be the normal  $p$ -complement of  $N$ . It is easy to see that  $N_{p'} \triangleleft G$ , so that  $N_{p'} = 1$  by Step 4 and  $N$  is a nontrivial  $p$ -group, contrary to Step 5. Consequently, we must have  $G = NP$ . From Step 6,  $G$  has Hall  $p'$ -subgroups. Then we may assume that  $N$  has a Hall  $p'$ -subgroup  $N_{p'}$ . By the Frattini argument,

$$G = NN_G(N_{p'}) = (P \cap N)N_{p'}N_G(N_{p'}) = (P \cap N)N_G(N_{p'}),$$

and thus

$$P = P \cap G = P \cap (P \cap N)N_G(N_{p'}) = (P \cap N)(P \cap N_G(N_{p'})).$$

Since  $N_G(N_{p'}) < G$ , we have  $P \cap N_G(N_{p'}) < P$ . We pick a maximal subgroup  $P_1$  of  $P$  such that  $P \cap N_G(N_{p'}) \leq P_1$ . Then  $P = (P \cap N)P_1$ . Let  $P_2$  be a 2-maximal subgroup of  $P$  such that  $P_2 \leq P_1$ . Under the hypothesis,  $P_2$  is either weakly  $s$ -permutably embedded or weakly  $s$ -supplemented in  $G$ . If  $P_2$  is weakly  $s$ -permutably embedded in  $G$ , then there is a subnormal subgroup  $T$  of  $G$  and an  $s$ -permutably embedded subgroup  $(P_2)_{se}$  of  $G$  contained in  $P_2$  such that  $G = P_2T$  and  $P_2 \cap T \leq (P_2)_{se}$ . Hence there is an  $s$ -permutable subgroup  $K$  of  $G$  such that  $(P_2)_{se}$  is a Sylow  $p$ -subgroup of  $K$ . If  $K_G \neq 1$ , then  $N \leq K_G \leq K$  and so  $(P_2)_{se} \cap N$  is a Sylow  $p$ -subgroup of  $N$ . We have that  $(P_2)_{se} \cap N \leq P_2 \cap N \leq P \cap N$  and  $P \cap N$  is a Sylow  $p$ -subgroup of  $N$ , thus  $(P_2)_{se} \cap N = P_2 \cap N = P \cap N$ . Consequently,

$$P = (N \cap P)P_1 = (P_2 \cap N)P_1 = P_1,$$

which is a contradiction. Thus  $K_G = 1$ . By Lemma 2.6,  $(P_2)_{se}$  is  $s$ -permutable in  $G$ . It follows from Lemma 2.11 that  $P_2 \cap T \leq (P_2)_{se} \leq O_p(G) = 1$ . Since  $|T|_p = p^2$ ,  $T$  is  $p$ -nilpotent by Lemma 2.4. Let  $T_{p'}$  be the normal  $p$ -complement of  $T$ . Then  $T_{p'}$  is a normal Hall  $p'$ -subgroup of  $G$ , a contradiction. Consequently, we may assume  $P_2$  is weakly  $s$ -supplemented in  $G$ . Then there is a subgroup  $T$  of  $G$  such that  $G = P_2T$  and  $P_2 \cap T \leq (P_2)_{sG} \leq O_p(G) = 1$  (where  $O_p(G)$  denotes the  $p$ -core of  $G$ ) by Lemma 2.11. Since  $|T|_p = p^2$ ,  $T$  is  $p$ -nilpotent by Lemma 2.4. Let  $T_{p'}$  be the normal  $p$ -complement of  $T$ . Then  $T_{p'}$  is a Hall  $p'$ -subgroup of  $G$ . By Step 6,

$T_{p'}$  and  $N_{p'}$  are conjugate in  $G$ . Since  $T_{p'}$  is normalized by  $T$ , there exists  $g \in P_2$  such that  $T_{p'}^g = N_{p'}$ . Hence

$$G = (P_2T)^g = P_2T^g = P_2N_G(T_{p'}^g) = P_2N_G(N_{p'})$$

and

$$P = P \cap G = P \cap P_2N_G(N_{p'}) = P_2(P \cap N_G(N_{p'})) \leq P_1,$$

a final contradiction. □

The following corollaries are immediate from Theorem 3.1.

**Corollary 3.2.** *Let  $p$  be the smallest prime dividing  $|G|$  and suppose  $G$  is  $A_4$ -free. Assume that  $H$  is a normal subgroup of  $G$  such that  $G/H$  is  $p$ -nilpotent. If there exists a Sylow  $p$ -subgroup  $P$  of  $H$  such that every 2-maximal subgroup of  $P$  is either weakly  $s$ -permutably embedded or weakly  $s$ -supplemented in  $G$ , then  $G$  is  $p$ -nilpotent.*

**Corollary 3.3.** *Suppose that every 2-maximal subgroup of any Sylow subgroup of a group  $G$  is either weakly  $s$ -permutably embedded or weakly  $s$ -supplemented in  $G$ . If  $G$  is  $A_4$ -free, then  $G$  is a Sylow tower group of supersolvable type.*

In terms of the theory of formations, we have the following result:

**Corollary 3.4.** *Let  $\mathcal{F}$  be the class of groups with Sylow tower of supersolvable type and suppose  $G$  is  $A_4$ -free. Then  $G \in \mathcal{F}$  if and only if there is a normal subgroup  $H$  of  $G$  such that  $G/H \in \mathcal{F}$  and every 2-maximal subgroup of any Sylow subgroup of  $H$  is either weakly  $s$ -permutably embedded or weakly  $s$ -supplemented in  $G$ .*

*Proof.* The necessity part is clear. We only need show the sufficiency part. Suppose that this is not true and let  $G$  be a counterexample of minimal order. By Lemmas 2.2 and 2.3, every 2-maximal subgroup of any Sylow subgroup of  $H$  is either weakly  $s$ -permutably embedded or weakly  $s$ -supplemented in  $H$ . By Corollary 3.3,  $H$  is a Sylow tower group of supersolvable type. Let  $p$  be the maximal prime divisor of  $|H|$  and let  $P$  be a Sylow  $p$ -subgroup of  $H$ . Then  $P$  is normal in  $G$ . Consider the factor group  $G/P$ . It is easy to prove  $G/P$  meets the hypotheses of the theorem. By the minimal choice of  $G$ , we get  $G/P \in \mathcal{F}$ . Let  $N$  be a minimal normal subgroup of  $G$  contained in  $P$ . The proof is divided into two steps.

Step 1:  $P = N$ .

If  $N < P$ , then  $(G/N)/(P/N) \cong G/P \in \mathcal{F}$ . We will prove that  $G/N \in \mathcal{F}$ . If  $|P/N| \leq p^2$ , then  $G/N \in \mathcal{F}$  by Lemma 2.4. If  $|P/N| > p^2$ , then every 2-maximal subgroup of  $P/N$  is either weakly  $s$ -permutably embedded or weakly  $s$ -supplemented in  $G/N$  by Lemmas 2.2 and 2.3. By the minimal choice of  $G$ , we get  $G/N \in \mathcal{F}$ . Since  $\mathcal{F}$  is a saturated formation,  $N$  is the unique minimal normal subgroup of  $G$  contained in  $P$  and  $N \not\leq \Phi(G)$ . It follows from Lemma 2.7 that  $P = F(P) = N$ , which is a contradiction.

Step 2: The final contradiction.

If  $|N| \leq p^2$ , then  $G \in \mathcal{F}$  by Lemma 2.9, a contradiction. Then  $|N| \geq p^3$ . Since  $N \triangleleft G$ , we may pick a 2-maximal subgroup  $N_2$  of  $N$  such that  $N_2 \triangleleft G_p$ , where  $G_p$  is a Sylow  $p$ -subgroup of  $G$ . Then  $N_2$  is either weakly  $s$ -permutably embedded or weakly  $s$ -supplemented in  $G$ . Let  $T$  be a supplement of  $N_2$  in  $G$ . Then  $G = N_2T = NT$  and  $N = N \cap N_2T = N_2(N \cap T)$ . This means that  $N \cap T \neq 1$ . However, since  $N \cap T$  is normal in  $G$  and  $N$  is minimal normal in  $G$ , we get  $N \cap T = N$  and thus  $T = G$ . If  $N_2$  is weakly  $s$ -permutably embedded in  $G$ , then  $(N_2)_{se} \geq N_2 \cap G = N_2$  is  $s$ -permutably embedded in  $G$ . From Lemma 2.6,  $N_2$  is  $s$ -permutable in  $G$  and  $O^p(G) \leq N_G(N_2)$ , where  $O^p(G)$  denotes the  $p$ -residual subgroup.<sup>1</sup> Thus  $N_2 \triangleleft G_p O^p(G) = G$ . It follows that  $|N| = p^2$ , a contradiction. If  $N_2$  is weakly  $s$ -supplemented in  $G$ , then  $N_2 = N_2 \cap G \leq (N_2)_{sG}$ . Similarly, we also get that  $N_2 \triangleleft G$ . We obtain the same contradiction, completing the proof.  $\square$

### Acknowledgements

The authors cordially thank the referee for valuable comments which led to the improvement of this paper.

### References

- [Ballester-Bolinches and Guo 1999] A. Ballester-Bolinches and X. Guo, “On complemented subgroups of finite groups”, *Arch. Math. (Basel)* **72** (1999), 161–166. MR 2000a:20037 Zbl 0929.20015
- [Ballester-Bolinches and Pedraza-Aguilera 1998] A. Ballester-Bolinches and M. C. Pedraza-Aguilera, “Sufficient conditions for supersolubility of finite groups”, *J. Pure Appl. Algebra* **127**:2 (1998), 113–118. MR 99d:20048 Zbl 0928.20020
- [Doerk and Hawkes 1992] K. Doerk and T. Hawkes, *Finite soluble groups*, de Gruyter Expositions in Mathematics **4**, Walter de Gruyter, Berlin, 1992. MR 93k:20033 Zbl 0753.20001
- [Gross 1987] F. Gross, “Conjugacy of odd order Hall subgroups”, *Bull. London Math. Soc.* **19**:4 (1987), 311–319. MR 89c:20038 Zbl 0616.20007
- [Guo and Shum 2001] X. Guo and K. P. Shum, “On  $c$ -normal subgroups of finite groups”, *Publ. Math. Debrecen* **58**:1-2 (2001), 85–92. MR 2001k:20034 Zbl 1062.20503
- [Guo and Shum 2003] X. Guo and K. P. Shum, “Cover-avoidance properties and the structure of finite groups”, *J. Pure Appl. Algebra* **181**:2-3 (2003), 297–308. MR 2004g:20027 Zbl 1028.20014
- [Guo et al. 2009] W. Guo, F. Xie, and B. Li, “Some open questions in the theory of generalized permutable subgroups”, *Sci. China Ser. A* **52**:10 (2009), 2132–2144. MR 2010k:20027 Zbl 1193.20021
- [Kegel 1962] O. H. Kegel, “Sylow-Gruppen und Subnormalteiler endlicher Gruppen”, *Math. Z.* **78** (1962), 205–221. MR 26 #5042 Zbl 0102.26802
- [Li and Guo 2000] D. Li and X. Guo, “The influence of  $c$ -normality of subgroups on the structure of finite groups”, *J. Pure Appl. Algebra* **150**:1 (2000), 53–60. MR 2001c:20032 Zbl 0967.20011

<sup>1</sup>This mean  $O^p(G)$  is the intersection of all normal subgroups of  $G$  whose index in  $G$  is a power of  $k$ . The quotient  $G/O^p(G)$  is the largest (not necessarily abelian)  $p$ -group onto which  $G$  surjects.

- [Li et al. 2005] Y. Li, Y. Wang, and H.-Q. Wei, “On  $p$ -nilpotency of finite groups with some subgroups  $\pi$ -quasinormally embedded”, *Acta Math. Hungar.* **108**:4 (2005), 283–298. MR 2006f:20022 Zbl 1094.20007
- [Li et al. 2009] Y. Li, S. Qiao, and Y. Wang, “On weakly  $s$ -permutably embedded subgroups of finite groups”, *Comm. Algebra* **37**:3 (2009), 1086–1097. MR 2010a:20041 Zbl 1177.20036
- [Robinson 1982] D. J. S. Robinson, *A course in the theory of groups*, Graduate Texts in Mathematics **80**, Springer, New York, 1982. 2nd ed. published in 1996. MR 84k:20001 Zbl 0483.20001
- [Schmid 1998] P. Schmid, “Subgroups permutable with all Sylow subgroups”, *J. Algebra* **207**:1 (1998), 285–293. MR 99g:20037 Zbl 0910.20015
- [Skiba 2007] A. N. Skiba, “On weakly  $s$ -permutable subgroups of finite groups”, *J. Algebra* **315**:1 (2007), 192–209. MR 2008k:20043 Zbl 1130.20019
- [Wang 1996] Y. Wang, “ $c$ -normality of groups and its properties”, *J. Algebra* **180**:3 (1996), 954–965. MR 97b:20020 Zbl 0847.20010
- [Wang 2000] Y. Wang, “Finite groups with some subgroups of Sylow subgroups  $c$ -supplemented”, *J. Algebra* **224**:2 (2000), 467–478. MR 2001c:20036 Zbl 0953.20010
- [Yang et al. 2012] N. Yang, W. Guo, J. Huang, and M. Xu, “Finite groups with weakly  $S$ -quasinormally embedded subgroups”, *J. Algebra Appl.* **11**:3 (2012), 1250050, 14. MR 2928118 Zbl 1244.20021
- [Zhang and Li 2012] X. Zhang and C. Li, “On weakly  $s$ -quasinormally embedded and  $c$ -supplemented subgroups of finite groups”, *Southeast Asian Bull. Math.* **36**:2 (2012), 293–300. MR 2992484 Zbl 06128975

Received: 2013-01-10      Revised: 2013-08-06      Accepted: 2013-08-07

zhg102003@163.com	<i>School of Mathematical Sciences, Guangxi Teachers Education University, Nanning, Guangxi 530023, China</i>
709725875@qq.com	<i>School of Mathematical Sciences, Beijing Normal University, Beijing 100875, China</i>
shixunlin@live.com	<i>College of Mathematics and Statistics, Zhaotong University, Zhaotong, Yunnan 657000, China</i>
305612276@qq.com	<i>School of Mathematical Sciences, Guangxi Teachers Education University, Nanning, Guangxi 530023, China</i>
407156835@qq.com	<i>School of Mathematical Sciences, Guangxi Teachers Education University, Nanning, Guangxi 530023, China</i>

# Ordering graphs in a normalized singular value measure

Charles R. Johnson, Brian Lins, Victor Luo and Sean Meehan

(Communicated by Joshua Cooper)

A proposed measure of network cohesion for graphs arising from interrelated economic activity is studied. The measure is the largest singular value of a row-stochastic matrix derived from the adjacency matrix. It is shown here that among graphs on  $n$  vertices, the star universally gives the (strictly) largest measure. Other universal comparisons among graphs with larger measures are difficult to make, but one is conjectured, and a selection of empirical evidence is given.

## 1. Introduction

In [Cavalcanti et al. 2012; 2013] the authors studied the role of network “cohesion” in the equilibration of economic or other activity among agents whose interaction is governed by a particular graph. An example is the one in which adjacency is the bordering relationship among countries. Giannitsarou and Johnson (personal communication, 2011) proposed a particular numerical measure of network cohesion and raised the question of which graph on  $n$  vertices resulted in the highest measure. That measure may be described as follows. Let  $A$  be the adjacency matrix of a graph  $G$ , define  $B = A + I$ , and let  $D$  be the positive diagonal matrix whose diagonal entries are the row sums of  $B$ . If  $R = D^{-1}B$ , then  $R$  is row-stochastic, and  $\sigma(G)$ , the measure of cohesion, is the largest singular value of  $R$ . Recall that the singular values of  $R$  are the square roots of the eigenvalues of  $RR^T$ . Another application where the matrix  $R$  has appeared is in [Echenique and Fryer 2007], where it is referred to as the matrix of social interactions.

Here, we show that, for any  $n$ ,  $\sigma(G)$  is maximized by the star  $S_n$ . The measure  $\sigma(G)$  is 1 if and only if  $G$  is regular, and 1 is the smallest possible value (Section 2, Proposition 1). Using our methods, it is difficult to determine, in advance, the relative position in this order of other graphs. Indeed, for graphs naturally defined on any number of vertices, the position often changes with  $n$ . However, we do

---

*MSC2010:* 05C40, 15A18.

*Keywords:* graph singular values, graph measure, network cohesion.

This work was partially supported by NSF grant DMS-0751964.

conjecture that the star plus an edge that connects two of the pendant vertices is next after the star, based, in part, on empirical evidence. After that, however, there may be no universal third place independent of  $n$ .

In the next section we mention known results that we use, and develop some new ideas that are important for our observations. In particular, the entries of  $RR^T$  have a nice and useful interpretation. Then, we show the star yields the highest measure by showing that a lower bound for the square of its largest singular value beats an upper bound for that of any other graph. Finally, in an Appendix, we give a selection of empirical information of interest (Table 1 and Figures 2, 3, 4, 5).

### 2. Background and tools

Given a graph  $G$  on  $n$  vertices, let  $A$  be the adjacency matrix of  $G$ . Unless otherwise noted, our notation follows [West 1996]. Let  $R = D^{-1}(A + I)$ , where  $D$  is the unique positive diagonal matrix such that  $R$  is row-stochastic. Let  $\lambda(G)$  denote the maximum eigenvalue of  $RR^T$ , and note that  $\sigma(G) = \sqrt{\lambda(G)}$ .

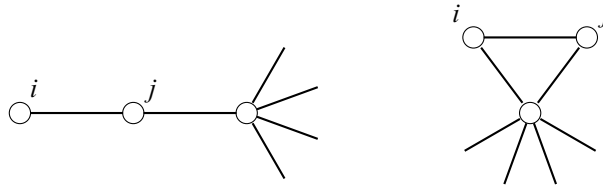
**Proposition 1.** *For any connected graph  $G$  on  $n$  vertices,  $\sigma(G) \geq 1$ , and  $\sigma(G) = 1$  if and only if  $G$  is regular.*

*Proof.* Note that  $G$  is regular if and only if  $R$  is doubly stochastic. If  $R$  is doubly stochastic, then it is a convex combination of permutation matrices by Birkhoff’s theorem [Horn and Johnson 1990, Theorem 8.1.7], and therefore the operator norm of  $R$ , which equals the maximum singular value, is 1. Let  $e \in \mathbb{R}^n$  denote the vector with 1 in every entry. By the Cauchy–Schwarz inequality,  $\|e^T R\|_2 \geq \langle e^T R, e/\sqrt{n} \rangle = \sqrt{n} = \|e^T\|_2$ , with equality if and only if  $e^T R$  is a multiple of  $e^T$ . Therefore, when  $R$  is row-stochastic but not doubly stochastic, the operator norm of  $R$  is strictly greater than one. It follows that  $\sigma(G) > 1$  when  $G$  is not regular.  $\square$

Note that  $D = \text{diag}(\{d_i + 1\}_{i \in 1, \dots, n})$ , where  $d_i$  is the degree of vertex  $i$  in  $G$ . Let  $C = (A + I)(A + I)^T$ . The  $(i, j)$  entry of  $C$ , which we denote by  $c_{ij}$ , is the number of vertices that are adjacent to both vertex  $i$  and vertex  $j$ , with the convention that two adjacent vertices are common neighbors of each other, that is,  $c_{ij} = |N[i] \cap N[j]|$ . In particular  $c_{ii} = d_i + 1$ . Thus the entries of  $RR^T$  are

$$r_{ij} = \frac{c_{ij}}{(d_i + 1)(d_j + 1)}. \tag{1}$$

**Lemma 1.** *Let  $RR^T$  be defined as above and assume that  $n > 2$ . When  $i \neq j$ , the largest possible values of  $r_{ij}$  are  $\frac{1}{3}$  and  $\frac{1}{4}$ . If  $r_{ij} = \frac{1}{3}$  for some  $i \neq j$ , then  $d_i = d_j = 2$  with  $c_{ij} = 3$  or  $\{d_i, d_j\} = \{1, 2\}$  with  $c_{ij} = 2$  (see Figure 1).*



**Figure 1.** Possible adjacency graphs when  $r_{ij} = \frac{1}{3}$ .

*Proof.* We may assume that  $d_j \geq d_i$ . Note that  $c_{ij} \leq d_i + 1$ ; thus  $r_{ij} \leq 1/(d_j + 1)$ . If  $r_{ij} > \frac{1}{4}$ , then  $d_j = 1$  or  $d_j = 2$ . In the former case,  $d_i = d_j = 1$ , which can only happen if  $n = 2$ , since  $G$  is assumed to be connected. In the latter case,  $d_i = 1$  or  $d_i = 2$  while  $d_j = 2$ . If  $d_i = 1$  and  $d_j = 2$ , then  $r_{ij} = c_{ij}/6 \in \{0, \frac{1}{6}, \frac{1}{3}\}$ , depending on the value of  $c_{ij}$ . If  $d_i = d_j = 2$ , then  $r_{ij} = c_{ij}/9 \in \{0, \frac{1}{9}, \frac{2}{9}, \frac{1}{3}\}$ .  $\square$

Suppose that  $G$  is a connected graph with  $n$  vertices such that every vertex has degree 1 (is pendant) except for a single central vertex with degree  $n - 1$ . We refer to any such graph as a *star* on  $n$  vertices, denoted by  $S_n$ . We may assume without loss of generality that vertex 1 is the central vertex of the star. Using (1), we see that, for the star,

$$RR^T = \begin{bmatrix} \frac{g}{n} & \frac{1}{n} & \dots & \dots & \frac{1}{n} \\ \frac{g}{n} & \frac{1}{2} & \frac{1}{4} & \dots & \frac{1}{4} \\ \vdots & \frac{1}{4} & \frac{1}{2} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \frac{1}{4} \\ \frac{g}{n} & \frac{1}{4} & \dots & \frac{1}{4} & \frac{1}{2} \end{bmatrix}.$$

Note that  $RR^T - \frac{1}{4}I$  is of rank 2, and therefore it is possible to explicitly calculate the characteristic polynomial of this matrix. Recall [Horn and Johnson 1990, Theorem 1.2.12] that the characteristic polynomial of a matrix is given by

$$p(t) = t^n - E_1 t^{n-1} + E_2 t^{n-2} + \dots + (-1)^n E_n,$$

where each  $E_k$  is the sum of the  $k$ -by- $k$  principal minors of the matrix. For  $RR^T - \frac{1}{4}I$ , only the 1-by-1 and 2-by-2 principal minors can be nonzero. Thus the characteristic equation for  $RR^T - \frac{1}{4}I$  is

$$p(t) = t^n - \left(\frac{1}{n} + \frac{1}{4}(n - 1)\right)t^{n-1} + \left(\frac{n - 4}{4n^2}\right)(n - 1)t^{n-2}.$$

The nonzero roots of this polynomial are

$$\frac{\frac{1}{4}(n-1) + \frac{1}{n} \pm \sqrt{\left(\frac{1}{4}(n-1) + \frac{1}{n}\right)^2 - \frac{n-4}{n^2}}}{2},$$

and therefore the maximum eigenvalue of  $RR^T$  for the star on  $n$  vertices is

$$\lambda(S_n) = \frac{1}{4} + \frac{\frac{1}{4}(n-1) + \frac{1}{n} + \sqrt{\left(\frac{1}{4}(n-1) + \frac{1}{n}\right)^2 - \frac{n-4}{n^2}}}{2}.$$

### 3. The star is a maximum

We seek to estimate the maximum eigenvalue  $\lambda(G)$  of  $RR^T$ . The row sums of  $RR^T$  place constraints on  $\lambda(G)$ . By [Horn and Johnson 1990, Theorem 8.1.22],

$$\min_i \left\{ \sum_j r_{ij} \right\} \leq \lambda(G) \leq \max_i \left\{ \sum_j r_{ij} \right\}. \tag{2}$$

For the star on  $n$  vertices,  $RR^T - \frac{1}{4}I$  contains an  $(n-1)$ -by- $(n-1)$  submatrix with all entries equal to  $\frac{1}{4}$ . It follows from the inclusion principle [Horn and Johnson 1990, Theorem 4.3.15] that  $\lambda(S_n) \geq \frac{1}{4}n$ . Combining this with the maximum row sum, we see that  $\frac{1}{4}n \leq \lambda(S_n) \leq \frac{1}{4}n + \frac{1}{n}$ .

The following observation is an immediate consequence of Lemma 1:

**Lemma 2.** *Suppose that  $n > 2$ , and consider the rows of  $RR^T$ . If row  $i$  has diagonal entry  $r_{ii} = \frac{1}{k}$  with  $k \geq 4$  and no off-diagonal entry equals  $\frac{1}{3}$ , then the sum of the entries in row  $i$  is at most  $\frac{1}{k} + \frac{1}{4}(n-1)$ .*

Let us make a basic observation which we will use in the proofs of several subsequent propositions.

**Lemma 3.** *Let  $c > 0$ . The function  $x \mapsto 1/(x+1) + cx$  is concave up for all  $x > 0$ , and therefore its maximum on any interval  $[a, b] \subset (0, \infty)$  is attained at one of the endpoints.*

The following observations about the row sums of  $RR^T$  cover the cases when Lemma 2 does not apply:

**Lemma 4.** *Suppose that  $n > 3$ . If row  $i$  has diagonal entry  $r_{ii} = \frac{1}{2}$  and  $G$  is not the star, then the sum of the entries in row  $i$  is at most  $-\frac{1}{6} + \frac{1}{n} + \frac{1}{4}n$ .*

*Proof.* Since  $r_{ii} = \frac{1}{2}$ ,  $d_i = 1$ . Let  $j$  denote the vertex adjacent to  $i$ . The sum of the entries in row  $i$  is then

$$r_{ii} + r_{ij} + \sum_{m \neq i, j} r_{im} = \frac{1}{2} + \frac{1}{d_j + 1} + \sum_{m \neq i, j} \frac{c_{im}}{d_m + 1}.$$



Note that  $c_{im} = 1$  if there is an edge connecting vertex  $j$  to vertex  $m$  and  $c_{im} = 0$  otherwise. Therefore we have the following upper bound for the sum of entries in row  $i$ :

$$r_{ii} + r_{ij} + \sum_{m \neq i, j} r_{im} \leq \frac{1}{2} + \frac{1}{d_j + 1} + \sum_{m \in N(j)} \frac{1}{2(d_m + 1)}.$$

If  $d_j = n - 1$ , and the graph is not the star, then there must be at least two vertices  $m_1$  and  $m_2$  such that  $d_{m_1} > 1$  and  $d_{m_2} > 1$ . In this case an upper bound for the sum of the entries in row  $i$  is

$$\frac{1}{2} + \frac{1}{n} + \frac{1}{4}(d_j - 3) + 2 \frac{1}{6} = -\frac{1}{6} + \frac{1}{n} + \frac{1}{4}n.$$

If  $d_j < n - 1$ , then

$$\begin{aligned} r_{ii} + r_{ij} + \sum_{m \neq i, j} r_{im} &\leq \frac{1}{2} + \frac{1}{d_j + 1} + \sum_{m \in N(j)} \frac{1}{2(d_m + 1)} \\ &\leq \frac{1}{2} + \frac{1}{d_j + 1} + \frac{1}{4}(d_j - 1). \end{aligned}$$

Since  $2 \leq d_j < n - 1$ , we use Lemma 3 to see that an upper bound for this expression is

$$\max\left\{\frac{13}{12}, -\frac{1}{4} + \frac{1}{n-1} + \frac{1}{4}n\right\}.$$

For  $n > 3$ ,

$$\max\left\{\frac{13}{12}, -\frac{1}{4} + \frac{1}{n-1} + \frac{1}{4}n\right\} \leq -\frac{1}{6} + \frac{1}{n} + \frac{1}{4}n. \quad \square$$

**Lemma 5.** *Suppose  $n > 3$ . If row  $i$  has diagonal entry  $r_{ii} = \frac{1}{3}$ , then the sum of the entries in row  $i$  is less than  $-\frac{1}{6} + \frac{1}{n} + \frac{1}{4}n$ .*

*Proof.* Since  $r_{ii} = \frac{1}{3}$ ,  $d_i = 2$ . Let  $j$  and  $k$  denote the two vertices adjacent to  $i$ .

*Case I.* If there is an edge connecting  $j$  and  $k$ , then  $c_{ij} = c_{ik} = 3$ . If  $m \neq i$  is a vertex adjacent to both  $j$  and  $k$ , then

$$r_{im} = \frac{2}{3(d_m + 1)} \leq \frac{2}{9}.$$

If  $m$  is only adjacent to one of  $j$  or  $k$ , then

$$r_{im} = \frac{1}{3(d_m + 1)} \leq \frac{1}{6}.$$

Let  $d = \max\{d_j, d_k\}$  and  $D = \max\{d_j, d_k\}$ . There are at most  $d - 2$  vertices other than  $i$  that are common neighbors of both  $j$  and  $k$ , and there are at most  $D - d$  remaining vertices other than  $i$  that could be adjacent to exactly one of  $j$  or  $k$ . Therefore the sum of the entries in row  $i$  is at most

$$r_{ii} + r_{ij} + r_{ik} + \frac{2}{9}(d - 2) + \frac{1}{6}(D - d) \leq -\frac{1}{9} + \frac{1}{(d + 1)} + \frac{1}{(D + 1)} + \frac{1}{18}d + \frac{1}{6}D.$$

In this case,  $2 \leq d \leq D \leq n - 1$ . By Lemma 3, it follows that the possible maximum values in the expression above occur when either  $d = D = 2$ , or  $d = 2, D = n - 1$ , or  $d = D = n - 1$ . The corresponding upper bounds on the row sum are

$$1, \quad \frac{1}{6} + \frac{1}{n} + \frac{1}{6}n, \quad -\frac{1}{3} + \frac{2}{n} + \frac{2}{9}n.$$

Each of these bounds is less than  $-\frac{1}{6} + \frac{1}{n} + \frac{1}{4}n$  for all  $n > 3$ .

*Case II.* If there is no edge connecting  $j$  with  $k$ , then  $c_{ij} = c_{ik} = 2$ . If  $m \neq i$  is a vertex adjacent to both  $j$  and  $k$ , then

$$r_{im} = \frac{2}{3(d_m + 1)} \leq \frac{2}{9}.$$

If  $m$  is only adjacent to one of  $j$  or  $k$ , then

$$r_{im} = \frac{1}{3(d_m + 1)} \leq \frac{1}{6}.$$

Let  $d = \max\{d_j, d_k\}$  and  $D = \max\{d_j, d_k\}$ . There are at most  $d - 1$  vertices other than  $i$  that are common neighbors of both  $j$  and  $k$ , and there are at most  $D - d$  remaining vertices other than  $i$  that could be adjacent to exactly one of  $j$  or  $k$ . Therefore the sum of the entries in row  $i$  is at most

$$r_{ii} + r_{ij} + r_{ik} + \frac{2}{9}(d - 1) + \frac{1}{6}(D - d) \leq \frac{1}{9} + \frac{2}{3(d + 1)} + \frac{2}{3(D + 1)} + \frac{1}{18}d + \frac{1}{6}D.$$

We know that  $1 \leq d \leq D \leq n - 2$ . By Lemma 3, it follows that the possible maximum values in the expression above occur when either  $d = D = 1$ , or  $d = 1, D = n - 2$ , or  $d = D = n - 2$ . The corresponding upper bounds on the row sum are

$$1, \quad \frac{1}{6} + \frac{2}{3(n - 1)} + \frac{1}{6}n, \quad -\frac{1}{3} + \frac{4}{3(n - 1)} + \frac{2}{9}n.$$

Once again, each of these bounds is less than  $-\frac{1}{6} + \frac{1}{n} + \frac{1}{4}n$  for all  $n > 3$ . □

**Lemma 6.** *Suppose  $n > 3$ . If row  $i$  contains an off-diagonal entry  $r_{ij} = \frac{1}{3}$ , then the sum of the entries in row  $i$  is at most  $-\frac{1}{6} + \frac{1}{n} + \frac{1}{4}n$ .*

*Proof.* There are three possible cases, depending on the possible degrees of  $i$  and  $j$  given by Lemma 1.

*Case I.* If  $d_i = 1$  and  $d_j = 2$ , then there is only one other vertex, aside from  $i$  and  $j$ , that can share a common neighbor with  $i$ . Call that vertex  $k$ . The sum of entries in row  $i$  is

$$r_{ii} + r_{ij} + r_{ik} = \frac{1}{2} + \frac{1}{3} + \frac{1}{2(d_k + 1)} \leq \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = \frac{13}{12},$$

which is less than or equal to  $-\frac{1}{6} + \frac{1}{n} + \frac{1}{4}n$  for all  $n > 3$  (equality occurs only when  $n = 4$ ).

*Case II.* If  $d_i = 2$  and  $d_j = 1$ , then Lemma 5 implies that the sum of the entries in row  $i$  is less than  $-\frac{1}{6} + \frac{1}{n} + \frac{1}{4}n$ .

*Case III.* If  $d_i = d_j = 2$ , then by Lemma 1,  $c_{ij} = 3$ . Let  $k$  denote the third common neighbor of  $i$  and  $j$ . The sum of the entries in row  $i$  is then

$$\begin{aligned} r_{ii} + r_{ij} + r_{ik} + \sum_{m \neq i, j, k} r_{im} &= \frac{1}{3} + \frac{1}{3} + \frac{1}{d_k + 1} + \sum_{m \neq i, j, k} \frac{1}{3(d_m + 1)} \\ &\leq \frac{2}{3} + \frac{1}{d_k + 1} + \frac{1}{6}(d_k - 2) \\ &\leq \frac{2}{3} + \frac{1}{n - 1} + \frac{1}{6}(n - 4) \\ &= \frac{1}{6}n + \frac{1}{n - 1}. \end{aligned}$$

This upper bound is less than  $-\frac{1}{6} + \frac{1}{n} + \frac{1}{4}n$  for all  $n > 3$ . □

**Theorem 1.** *Of all connected graphs on  $n$  vertices, the star attains the maximum value of  $\sigma$ .*

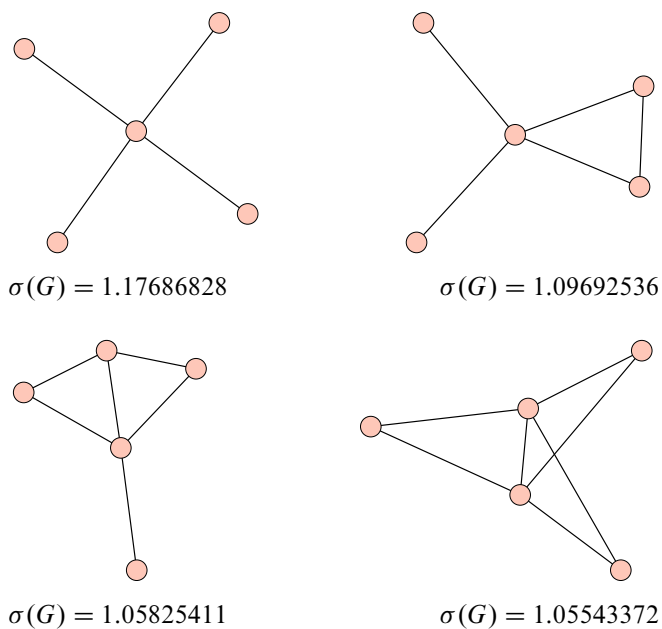
*Proof.* Suppose that  $G$  is not  $S_n$ . The contents of Lemmas 2, 4, 5, and 6 show that the maximum row sum of  $RR^T$  is less than or equal to  $-\frac{1}{6} + \frac{1}{n} + \frac{1}{4}n$ . If  $n > 6$ , then this upper bound is less than  $\frac{1}{4}n$ , and, by the comment after (2), we conclude that  $\lambda(G) < \lambda(S_n)$  and therefore  $\sigma(G) < \sigma(S_n)$ . When  $3 < n \leq 6$ , we can verify by explicit computation that  $-\frac{1}{6} + \frac{1}{n} + \frac{1}{4}n < \lambda(S_n)$ . When  $n = 3$ , the theorem can be verified directly since there are only two connected graphs on 3 vertices. □

### Appendix

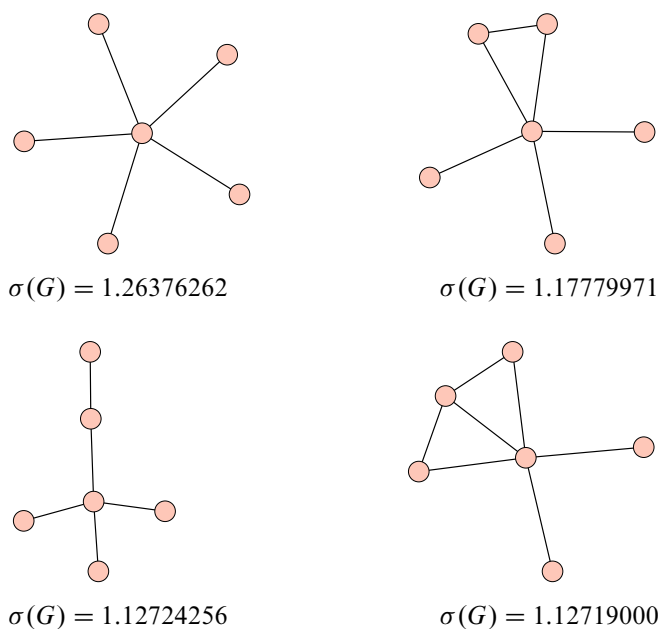
Here we present the values of  $\sigma(G)$  for every connected graph up to 6 vertices. The graphs are given in graph6 string format [McKay 1981; 2005], and the values of  $\sigma(G)$  are given to 8 decimal places. The values for the stars are given in boldface.

Esa?	<b>1.26376262</b>	EvsW	1.04127270	Elo_	1.02301009	Er_G	1.01059866
Eta?	1.17779971	Et}G	1.04082858	ExoG	1.02253862	Dxc	1.00995156
Ds_	<b>1.17686828</b>	Ev{W	1.04057352	E~{W	1.02245280	EpOG	1.00969514
Epa?	1.12724256	EzPW	1.03944703	EvwW	1.02150256	E~wW	1.00956370
Exg_	1.12719000	Elw_	1.03869527	E~sW	1.02136937	ExOW	1.00891795
E i_	1.11535507	EvcG	1.03802560	EzZW	1.02039571	ErOW	1.00885018
E g_	1.10702341	Dx_	1.03794998	EzoG	1.02034616	Ez_G	1.00805939
ExGg	1.10124485	Epo_	1.03760887	D~c	1.02031933	D~s	1.00764077
Dt_	1.09692536	Eto_	1.03627677	ErcG	1.01998619	EzYW	1.00741994
Cs	<b>1.09445053</b>	Dto	1.03552399	EzWW	1.01866302	ErOW	1.00711468
Exw_	1.09118881	ExPw	1.03508808	EzOW	1.01862583	Epoo	1.00711468
Ehg_	1.08965849	Exwo	1.03458078	EpUG	1.01823188	E~yW	1.00707898
Ex__	1.08492159	EtUG	1.03375811	Ez{W	1.01792742	ExSW	1.00696806
Eli_	1.08378641	Edq_	1.03272839	E~sG	1.01775521	ErWW	1.00696806
E __	1.08125829	EzZw	1.03266215	E qW	1.01732826	E oW	1.00662172
Ep{G	1.07743057	EpgG	1.03266215	Exoo	1.01710090	Ezsw	1.00608114
Elg_	1.07386856	Er{W	1.03197929	Ez{w	1.01709947	Ezow	1.00603467
Et}G	1.06680419	EzwG	1.03138546	EzSW	1.01709947	Dzs	1.00499991
E w_	1.06420788	Cx	1.03138184	Dxo	1.01695288	Dzc	1.00459536
Etq_	1.06264937	Ev_G	1.03126091	Ezww	1.01494232	E~}w	1.00451397
Exo_	1.06170523	Dxw	1.02998084	E~OW	1.01436311	E~uw	1.00445419
Ep__	1.06066017	EpWG	1.02979441	Cz	1.01417394	E OW	1.00293400
EtuG	1.05968917	E~TW	1.02813174	Cp	1.01417394	E~YW	1.00274201
ExGG	1.05861770	Bo	<b>1.02813174</b>	E sW	1.01400371	E~~w	1.00000000
D _	1.05825411	Ep_G	1.02808843	EroG	1.01390539	Ezuw	1.00000000
D g	1.05543372	ErwG	1.02792587	E~cG	1.01337635	Erow	1.00000000
Dp_	1.05417745	E~{G	1.02768976	ErwW	1.01293228	ErYW	1.00000000
Eh__	1.05150374	Dl_	1.02717603	E~}W	1.01273126	EpOW	1.00000000
El__	1.04879365	E~SW	1.02636956	Edo_	1.01267470	D~{	1.00000000
Er{G	1.04866795	EzsG	1.02530775	Dh_	1.01213081	Dhc	1.00000000
EpsG	1.04851433	Dlg	1.02465677	Dpo	1.01188403	C~	1.00000000
E o_	1.04562708	EvoW	1.02459474	E SW	1.01133377	Cr	1.00000000
ExwG	1.04512215	ExOG	1.02421645	E~_G	1.01111110	Bw	1.00000000
ExwG	1.04350178	ExPW	1.02380968	E TW	1.01090626	A_	<b>1.00000000</b>
EpuG	1.04308838	D c	1.02305146	EzcG	1.01084213		
Ez{G	1.04248210	EpSG	1.02303779	E~oW	1.01073140		

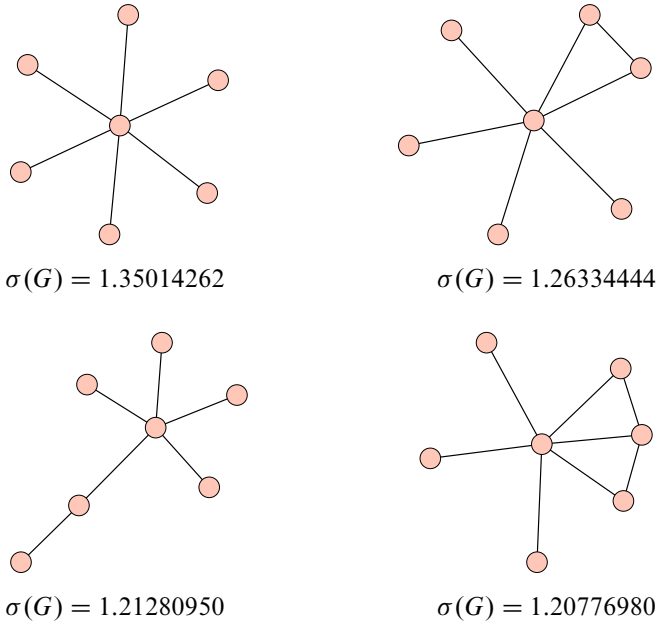
**Table 1.** The value of  $\sigma(G)$  (to 8 decimal places) for every connected graph with at most 6 vertices, with the values of stars given in boldface.



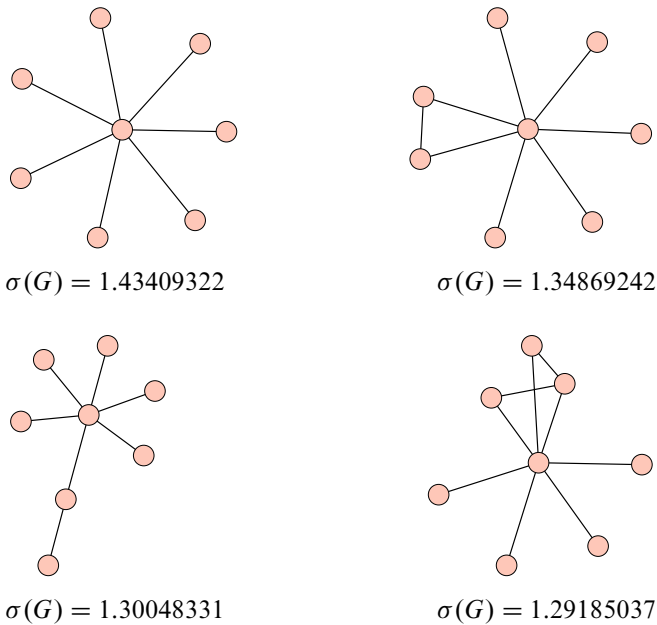
**Figure 2.** The graphs with the four highest singular values for  $n = 5$ .



**Figure 3.** The graphs with the four highest singular values for  $n = 6$ .



**Figure 4.** The graphs with the four highest singular values for  $n = 7$ .



**Figure 5.** The graphs with the four highest singular values for  $n = 8$ .

## References

- [Cavalcanti and Giannitsarou 2013] T. V. V. Cavalcanti and C. Giannitsarou, “Growth and human capital: a network approach”, April 23, 2013, available at [http://www.econ.cam.ac.uk/faculty/giannitsarou/net\\_growth.pdf](http://www.econ.cam.ac.uk/faculty/giannitsarou/net_growth.pdf). Submitted. Formerly titled “Network structure and human capital dynamics”.
- [Cavalcanti et al. 2012] T. V. V. Cavalcanti, C. Giannitsarou, and C. R. Johnson, “Network cohesion”, 2012, available at <http://www.econ.cam.ac.uk/faculty/giannitsarou/cohesion.pdf>.
- [Echenique and Fryer 2007] F. Echenique and R. G. Fryer, “A measure of segregation based on social interactions”, *Q. J. Econ.* **122**:2 (2007), 441–485.
- [Horn and Johnson 1990] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, 1990. Corrected reprint of the 1985 original. MR 91i:15001 Zbl 0704.15002
- [McKay 1981] B. D. McKay, “Practical graph isomorphism”, pp. 45–87 in *Proceedings of the Tenth Manitoba Conference on Numerical Mathematics and Computing* (Winnipeg, MB, 1980), edited by D. S. Meek and H. C. Williams, Congr. Numer. **30**, Utilitas Mathematical Publishing, Winnipeg, MB, 1981. MR 83e:05061 Zbl 0521.05061
- [McKay 2005] B. D. McKay, “Description of graph6 and sparse6 encodings”, 2005, available at <http://cs.anu.edu.au/~bdm/data/formats.txt>.
- [West 1996] D. B. West, *Introduction to graph theory*, Prentice Hall, Upper Saddle River, NJ, 1996. MR 96i:05001 Zbl 0845.05001

Received: 2013-03-06    Revised: 2013-07-18    Accepted: 2013-07-24

crjohnso@math.wm.edu    *Department of Mathematics, College of William & Mary,  
Williamsburg, VA 23187, United States*

blins@hsc.edu    *Mathematics and Computer Science Department,  
Hampden-Sydney College, Hampden Sydney, VA 23943,  
United States*

vd11@williams.edu    *Department of Mathematics and Statistics,  
Williams College, 2899 Paresky, 39 Chapin Hall Drive,  
Williamstown, MA 01267, United States*

smeehan1@nd.edu    *Department of Mathematics, University of Notre Dame,  
Notre Dame, IN 46556, United States*





# More explicit formulas for Bernoulli and Euler numbers

Francesca Romano

(Communicated by Ken Ono)

By directly considering Taylor coefficients and composite generating functions, we employ a generalized Faà di Bruno formula for higher partial derivatives using vector partitions to obtain identities that include explicit formulas for the Bernoulli and Euler numbers. The formulas we obtain are generalized analogs of the formulas obtained by D. C. Vella.

## 1. Introduction

The purpose of this paper is to extend the results of Vella [2008] using vector partitions. Recall that the sequences of Bernoulli numbers  $B_n$  and Euler numbers  $E_n$  have exponential generating functions  $x/(e^x - 1)$  and  $\operatorname{sech} x$  respectively. Vella obtained the identities

$$B_n = \sum_{\pi \in \mathcal{P}_n} \frac{(-1)^m}{1+m} \binom{m}{\lambda(\pi)} \binom{n}{\pi} = \sum_{\pi \in \mathcal{C}_n} \frac{(-1)^m}{1+m} \binom{n}{\pi},$$

$$B_n = \sum_{1 \leq m \leq n} \frac{(-1)^m m!}{1+m} S(n, m),$$

$$E_n = \sum_{\substack{\pi \in \mathcal{P}_n \\ \text{even parts}}} (-1)^m \binom{m}{\lambda(\pi)} \binom{n}{\pi} = \sum_{\substack{\pi \in \mathcal{C}_n \\ \text{even parts}}} (-1)^m \binom{n}{\pi},$$

$$E_n = \sum_{1 \leq m \leq n} (-1)^m m! S(n, m, \text{even}),$$

$$1 = \sum_{1 \leq r \leq j} \frac{(-1)^r}{(2r)!} E_{2r} \sum_{\substack{\pi \in \mathcal{P}_{2j, 2r} \\ \text{odd parts}}} \binom{2r}{\lambda(\pi)} \binom{2j}{\pi} \prod_{s=0}^j [E_{2s}]^{\pi_{2s+1}} \quad \text{for all } j > 0,$$

*MSC2010:* primary 11B68; secondary 05A15.

*Keywords:* Bernoulli numbers, Euler numbers, multivariable calculus.

where  $\mathcal{P}_n$  is the set of integer partitions of  $n$ ,  $\mathcal{C}_n$  is the set of all ordered partitions (i.e., compositions) of  $n$ ,  $m$  is the length of  $\pi$ ,  $\lambda(\pi)$  is the multiset of multiplicities of  $\pi$ ,  $S(n, m)$  is the Stirling number of the second kind, that is, the number of ways of partitioning a set of  $n$  elements into exactly  $m$  nonempty subsets, and  $S(n, m, \text{even})$  is the number of ways of partitioning a set of  $n$  elements into exactly  $m$  nonempty subsets each with even cardinality.

Let

$$h_B(x_1, \dots, x_\nu) = \frac{x_1 + \dots + x_\nu}{e^{x_1 + \dots + x_\nu} - 1} \quad \text{and} \quad h_E(x_1, \dots, x_\nu) = \text{sech}(x_1 + \dots + x_\nu)$$

be functions from  $\mathbb{R}^\nu$  into  $\mathbb{R}$ , where  $\nu \in \mathbb{N}$ . For a multiindex  $\alpha = (\alpha_1, \dots, \alpha_\nu) \in \mathbb{N}_0^\nu$  we consider the generalized Bernoulli number  $B_\alpha$  to be  $\alpha!$  times the  $\alpha$ -th Taylor coefficient of  $h_B$ . We define generalized Euler numbers analogously. These generalized Bernoulli numbers and Euler numbers were recently introduced and studied in [Di Nardo and Oliva 2012] in connection with multivariable Lévy processes. Note that although it wasn't explicitly said in [Di Nardo and Oliva 2012],  $B_\alpha = B_{|\alpha|}$ , where  $|\alpha| = \sum_{k=1}^\nu \alpha_k$ , and thus  $B_\alpha$  is simply the  $|\alpha|$ -th Bernoulli number. Also notice that if  $\alpha, \beta \in \mathbb{N}_0^\nu$  and  $|\alpha| = |\beta|$  then  $B_\alpha = B_\beta$ . The same can also be said for the Euler numbers. In the present paper, we prove the precise analogues of the identities above for these Bernoulli and Euler numbers by applying the multivariable Faà di Bruno formula found in [Constantine and Savits 1996].

The point of view adopted in [Vella 2008] is that thinking explicitly about Taylor coefficients yields tools with a lot of combinatorial leverage. The results of the present paper rely even more heavily on this point of view. For example, it would be interesting to have a combinatorial interpretation for the analogue of  $S(n, m)$  that appears in our new formulas, but we obtain these formulas without such a combinatorial interpretation.

## 2. Notation and review of vector partitions

In this section, we fix notation that parallels that used in [Constantine and Savits 1996] but will in the end yield formulas looking like those in [Vella 2008]. We also restate the results from [Constantine and Savits 1996] in our notation. Below let  $\mathbb{N}$  denote the set of natural numbers,  $\mathbb{N}_0$  the set of nonnegative integers. We regard finite cartesian powers, such as  $\mathbb{N}_0^\nu$ , and  $\mathbb{N}^\nu$ , where  $\nu \in \mathbb{N}$ , as sitting in the natural way in the real vector space  $\mathbb{R}^\nu$  throughout.

Since the generalized Faà di Bruno formula found in [Constantine and Savits 1996] is expressed as a sum over the vector partitions of  $\alpha = (\alpha_1, \dots, \alpha_\nu) \in \mathbb{N}_0^\nu$ , we begin with a review of the vector partition notation we have adopted in this paper. A *vector partition*  $\pi = (\mathbf{m}_1, \dots, \mathbf{m}_s; \mathbf{p}_1, \dots, \mathbf{p}_s)$  of  $\alpha$  is a multiset of *vector parts*  $\mathbf{p}_1, \dots, \mathbf{p}_s \in \mathbb{N}_0^\nu$  and their respective *vector multiplicities*  $\mathbf{m}_1, \dots, \mathbf{m}_s \in \mathbb{N}_0^\mu$  with

$\mu, s \in \mathbb{N}$ , where

$$\sum_{i=1}^s \mathbf{m}_i = \mathbf{m} = (r_1, \dots, r_\mu) \in \mathbb{N}_0^\mu, \quad |\mathbf{m}_i| = \sum_{j=1}^\mu m_{ij} > 0, \quad \sum_{i=1}^s |\mathbf{m}_i| \mathbf{p}_i = \boldsymbol{\alpha}.$$

Additionally, we require that the parts are lexicographically ordered, that is,

$$\mathbf{0} < \mathbf{p}_1 < \dots < \mathbf{p}_s,$$

where  $\mathbf{p}_i < \mathbf{p}_j$  means  $\mathbf{p}_i$  and  $\mathbf{p}_j$  satisfy one of the following:

- $|\mathbf{p}_i| < |\mathbf{p}_j|$ .
- $|\mathbf{p}_i| = |\mathbf{p}_j|$  and  $p_{i1} < p_{j1}$ .
- $|\mathbf{p}_i| = |\mathbf{p}_j|$  and  $p_{i1} = p_{j1}, p_{i2} = p_{j2}, \dots, p_{ik} = p_{jk}$  and  $p_{i(k+1)} < p_{j(k+1)}$  for some  $1 \leq k < v$ .

One readily checks that  $<$  defines a total ordering on  $\mathbb{N}_0^v$ .

The set of vector partitions of  $\boldsymbol{\alpha}$  of size  $s$  and total multiplicity  $\mathbf{m}$  is denoted by  $p_s(\boldsymbol{\alpha}, \mathbf{m})$ . Note that the size  $s$  is the number of vector parts in the partition and this number differs from the total multiplicity of the partition. We let

$$p(\boldsymbol{\alpha}, \mathbf{m}) = \bigcup_{s=1}^{|\boldsymbol{\alpha}|} p_s(\boldsymbol{\alpha}, \mathbf{m}) \quad \text{and} \quad p(\boldsymbol{\alpha}) = \bigcup_{1 \leq |\mathbf{m}| \leq |\boldsymbol{\alpha}|} p(\boldsymbol{\alpha}, \mathbf{m}),$$

and we always set  $0^0 = 1$ . The above definitions can be clarified by working through the example below.

**Example 2.1.** Let  $v = 3$  and  $\mu = 2$ . Take  $\boldsymbol{\alpha} = (1, 2, 1)$  where  $|\boldsymbol{\alpha}| = 4$ . We will verify that

$$\pi = ((1, 1), (1, 0); (0, 1, 0), (1, 0, 1)) \in p_2(\boldsymbol{\alpha}, \mathbf{m}),$$

where  $\mathbf{m} = (2, 1)$ . First observe that  $\sum_{i=1}^2 \mathbf{m}_i = (1, 1) + (1, 0) = (2, 1) = \mathbf{m}$  and  $|\mathbf{m}_1| = 2 > 0$  and  $|\mathbf{m}_2| = 1 > 0$ . Now observe that

$$\sum_{i=1}^2 |\mathbf{m}_i| \mathbf{p}_i = 2(0, 1, 0) + 1(1, 0, 1) = (1, 2, 1) = \boldsymbol{\alpha}.$$

Finally, our last condition is met because  $\mathbf{p}_1 < \mathbf{p}_2$  since  $|\mathbf{p}_1| = 1 < 2 = |\mathbf{p}_2|$ .

We will also make use of ordered vector partitions of  $\boldsymbol{\alpha}$ . The set of ordered vector partitions of  $\boldsymbol{\alpha}$  of total multiplicity  $\mathbf{m}$  is denoted by  $s^+(\boldsymbol{\alpha}, \mathbf{m})$ . In order to define  $s^+(\boldsymbol{\alpha}, \mathbf{m})$ , we must first define the following:

$$s(\boldsymbol{\alpha}, \mathbf{m}) = \left\{ (\mathbf{p}_1^{(1)}, \dots, \mathbf{p}_{r_1}^{(1)}; \dots; \mathbf{p}_1^{(\mu)}, \dots, \mathbf{p}_{r_\mu}^{(\mu)}): \mathbf{p}_j^{(i)} \in \mathbb{N}_0^v \text{ and } \sum_{i=1}^\mu \sum_{j=1}^{r_i} \mathbf{p}_j^{(i)} = \boldsymbol{\alpha} \right\}.$$

This allows us to define our set of ordered vector partitions as follows:

$$s^+(\alpha, \mathbf{m}) = \{(\mathbf{p}_1^{(1)}, \dots, \mathbf{p}_{r_1}^{(1)}; \dots; \mathbf{p}_1^{(\mu)}, \dots, \mathbf{p}_{r_\mu}^{(\mu)}) \in s(\alpha, \mathbf{m}) : \mathbf{p}_j^{(i)} \neq \mathbf{0}, i \in \{1, \dots, \mu\}, j \in \{1, \dots, r_i\}\}.$$

We let  $s^+(\alpha) = \bigcup_{1 \leq |\mathbf{m}| \leq |\alpha|} s^+(\alpha, \mathbf{m})$ . The definition of an ordered vector partition of  $\alpha$  of total multiplicity  $\mathbf{m}$  can be clarified by working through the example below.

**Example 2.2.** Take  $\nu = 3$  and  $\mu = 2$ , as before, with  $\alpha = (1, 2, 1)$ . We will first verify that  $\pi = ((0, 1, 0), (1, 0, 1); (0, 1, 0), (0, 0, 0)) \in s(\alpha, \mathbf{m})$  where  $\mathbf{m} = (2, 1)$ . Notice that the size of this ordered partition is 4, but  $|\mathbf{m}| = 3$ . Now observe that  $\sum_{i=1}^2 \sum_{j=1}^{r_i} \mathbf{p}_j^{(i)} = (0, 1, 0) + (1, 0, 1) + (0, 1, 0) + (0, 0, 0) = (1, 2, 1) = \alpha$ . Now we can construct an element  $\pi' \in s^+(\alpha, \mathbf{m})$  by removing all elements of  $\pi$  equal to  $(0, 0, 0)$ . Thus  $\pi' = ((0, 1, 0), (1, 0, 1); (0, 1, 0)) \in s^+(\alpha, \mathbf{m})$ . Notice that  $\pi$  is a different element of  $s(\alpha, \mathbf{m})$  than  $\pi'' = ((1, 0, 1), (0, 1, 0); (0, 1, 0), (0, 0, 0))$  and yields an element of  $s^+(\alpha, \mathbf{m})$  not equal to  $\pi'$ .

### 3. The generalized Faà di Bruno formula

We begin this section by restating the multiindex notation found on page 504 in [Constantine and Savits 1996], which will be used in the generalized Faà di Bruno formula. In what follows, let  $\alpha = (\alpha_1, \dots, \alpha_\nu) \in \mathbb{N}_0^\nu$ ,  $\mathbf{x} = (x_1, \dots, x_\nu) \in \mathbb{R}^\nu$  and

$$\alpha! = \prod_{i=1}^{\nu} (\alpha_i!), \quad \mathbf{x}^\alpha = \prod_{i=1}^{\nu} x_i^{\alpha_i},$$

$$D_{\mathbf{x}}^{\mathbf{0}} = \text{identity operator}, \quad D_{\mathbf{x}}^{\alpha} = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_\nu^{\alpha_\nu}} \quad \text{for } |\alpha| > 0.$$

Note that for  $\mathbf{w} = (w_1, \dots, w_\nu) \in \mathbb{N}_0^\nu$ , we write  $\mathbf{w} \leq \alpha$  if  $w_k \leq \alpha_k$  for  $k = 1, 2, \dots, \nu$ . A function  $h$  is an element of  $C_\alpha(\mathbf{x}^0)$  if  $D_{\mathbf{x}}^{\mathbf{w}} h$  exists and is continuous in a neighborhood of  $\mathbf{x}^0$  for all  $\mathbf{w} \leq \alpha$ . Additionally, a function  $h$  is an element of  $C^n(\mathbf{x}^0)$  if  $h \in C_{\mathbf{w}}(\mathbf{x}^0)$  for all  $|\mathbf{w}| \leq n$ .

Now let  $g : \mathbb{R}^\nu \rightarrow \mathbb{R}^\mu$  and  $f : \mathbb{R}^\mu \rightarrow \mathbb{R}$  be functions and  $h : \mathbb{R}^\nu \rightarrow \mathbb{R}$  their composition; that is, let

$$h(x_1, \dots, x_\nu) = f[g^{(1)}(x_1, \dots, x_\nu), \dots, g^{(\mu)}(x_1, \dots, x_\nu)].$$

Assume that  $\mathbf{0} \neq \alpha = (\alpha_1, \dots, \alpha_\nu) \in \mathbb{N}_0^\nu$  and  $\mathbf{x}^0 = (x_1^0, \dots, x_\nu^0) \in \mathbb{R}^\nu$  are given,  $g^{(1)}, \dots, g^{(\mu)} \in C_\alpha(\mathbf{x}^0)$  and  $f \in C^{|\alpha|}(\mathbf{y}^0)$ , where  $\mathbf{y}^0 = (g^{(1)}(\mathbf{x}^0), \dots, g^{(\mu)}(\mathbf{x}^0))$ . Then, setting  $h_\alpha = D_{\mathbf{x}}^\alpha h(\mathbf{x}^0)$ ,  $f_{\mathbf{m}} = D_{\mathbf{y}}^{\mathbf{m}} f(\mathbf{y}^0)$ ,  $g_{\mathbf{k}}^{(i)} = D_{\mathbf{x}}^{\mathbf{k}} g^{(i)}(\mathbf{x}^0)$ , and  $\mathbf{g}_{\mathbf{k}} = (g_{\mathbf{k}}^{(1)}, \dots, g_{\mathbf{k}}^{(\mu)})$ , we can state the generalized Faà di Bruno formula that appears as the main result (Theorem 2.1) of [Constantine and Savits 1996]:

**Theorem 3.1.** 
$$h_{\alpha} = \sum_{1 \leq |m| \leq |\alpha|} f_m \sum_{s=1}^{|\alpha|} \sum_{\pi \in p_s(\alpha, m)} (\alpha!) \prod_{j=1}^s \frac{[g_{p_j}]^{m_j}}{(m_j!)[p_j!]^{|m_j|}}.$$

The proof of the above theorem found in [Constantine and Savits 1996] takes into account issues of convergence. Now we can rigorously rewrite this generalized formula to resemble the single variable formula used in [Vella 2008]. First let

$$\binom{\alpha}{\pi} = \frac{\alpha!}{\pi!}, \quad \pi! = \prod_{j=1}^s [p_j!]^{|m_j|} \quad \text{and} \quad \lambda(\pi)! = \prod_{j=1}^s (m_j!).$$

Now observe,

$$\begin{aligned} h_{\alpha} &= \sum_{1 \leq |m| \leq |\alpha|} f_m \sum_{s=1}^{|\alpha|} \sum_{\pi \in p_s(\alpha, m)} (\alpha!) \prod_{j=1}^s \frac{[g_{p_j}]^{m_j}}{(m_j!)[p_j!]^{|m_j|}} \\ &= \sum_{1 \leq |m| \leq |\alpha|} (\alpha!) f_m \sum_{\pi \in p(\alpha, m)} \prod_{j=1}^s \frac{[g_{p_j}]^{m_j}}{(m_j!)[p_j!]^{|m_j|}} \\ &= \sum_{\pi \in p(\alpha)} \frac{\alpha!}{\prod_{j=1}^s (m_j!)[p_j!]^{|m_j|}} f_m \prod_{j=1}^s [g_{p_j}]^{m_j} \\ &= \sum_{\pi \in p(\alpha)} \frac{\binom{\alpha}{\pi}}{\lambda(\pi)!} f_m \prod_{j=1}^s [g_{p_j}]^{m_j}. \end{aligned} \tag{1}$$

Our formula for Taylor coefficients of  $h_{\alpha}$  follows:

**Corollary 3.2.**

$$T_{\alpha}(h; \mathbf{x}^0) = \sum_{1 \leq |m| \leq |\alpha|} T_m(f; \mathbf{y}^0) \sum_{\pi \in p(\alpha, m)} \binom{m}{\lambda(\pi)} \prod_{j=1}^s \prod_{k=1}^{\mu} [T_{p_j}(g^{(k)}; \mathbf{x}^0)]^{(m_j)_k}. \tag{2}$$

*Proof.* This follows directly from (1), since

$$\begin{aligned} T_{\alpha}(h; \mathbf{x}^0) &= \frac{h_{\alpha}}{\alpha!} = \sum_{\pi \in p(\alpha)} \frac{f_m}{\pi! \lambda(\pi)!} \prod_{j=1}^s [g_{p_j}]^{m_j} \\ &= \sum_{\pi \in p(\alpha)} \frac{m! f_m}{m! \lambda(\pi)!} \prod_{j=1}^s \frac{[g_{p_j}]^{m_j}}{[p_j!]^{|m_j|}} \\ &= \sum_{\pi \in p(\alpha)} \binom{m}{\lambda(\pi)} \frac{f_m}{m!} \prod_{j=1}^s \frac{\prod_{k=1}^{\mu} [g_{p_j}]^{(m_j)_k}}{[p_j!]^{\sum_{k=1}^{\mu} (m_j)_k}} \end{aligned}$$

$$\begin{aligned}
 &= \sum_{\pi \in p(\alpha)} \binom{\mathbf{m}}{\lambda(\pi)} \frac{f_{\mathbf{m}}}{\mathbf{m}!} \prod_{j=1}^s \prod_{k=1}^{\mu} \left( \frac{g_{\mathbf{p}_j}^{(k)}}{\mathbf{p}_j!} \right)^{(m_j)_k} \\
 &= \sum_{1 \leq |\mathbf{m}| \leq |\alpha|} \frac{f_{\mathbf{m}}}{\mathbf{m}!} \sum_{\pi \in p(\alpha, \mathbf{m})} \binom{\mathbf{m}}{\lambda(\pi)} \prod_{j=1}^s \prod_{k=1}^{\mu} \left( \frac{g_{\mathbf{p}_j}^{(k)}}{\mathbf{p}_j!} \right)^{(m_j)_k} \\
 &= \sum_{1 \leq |\mathbf{m}| \leq |\alpha|} T_{\mathbf{m}}(f; \mathbf{y}^0) \sum_{\pi \in p(\alpha, \mathbf{m})} \binom{\mathbf{m}}{\lambda(\pi)} \prod_{j=1}^s \prod_{k=1}^{\mu} [T_{\mathbf{p}_j}(g^{(k)}; \mathbf{x}^0)]^{(m_j)_k}. \quad \square
 \end{aligned}$$

We will also want to make use of the generalized Faà di Bruno formula that considered ordered vector partitions. This is given by Theorem 3.4 of [Constantine and Savits 1996]:

**Theorem 3.3.** 
$$h_{\alpha} = \alpha! \sum_{1 \leq |\mathbf{m}|} \frac{f_{\mathbf{m}}}{\mathbf{m}!} \sum_{\pi \in s(\alpha, \mathbf{m})} \prod_{i=1}^{\mu} \prod_{j=1}^{r_i} \frac{[g_{\mathbf{p}_j}^{(i)}]}{[\mathbf{p}_j^{(i)}!]}.$$
 (3)

**Proposition 3.4.** 
$$T_{\alpha}(h; \mathbf{x}^0) = \sum_{s^+(\alpha)} T_{\mathbf{m}}(f; \mathbf{y}^0) \prod_{i=1}^{\mu} \prod_{j=1}^{r_i} T_{\mathbf{p}_j^{(i)}}(g^{(i)}; \mathbf{x}^0).$$

*Proof.* This follows directly from (3) by substituting formulas 3.3 and 3.8 of [Constantine and Savits 1996] as follows:

$$\begin{aligned}
 T_{\alpha}(h; \mathbf{x}^0) &= \frac{h_{\alpha}}{\alpha!} = \sum_{1 \leq |\mathbf{m}|} \frac{f_{\mathbf{m}}}{\mathbf{m}!} \sum_{\pi \in s(\alpha, \mathbf{m})} \prod_{i=1}^{\mu} \prod_{j=1}^{r_i} \frac{[g_{\mathbf{p}_j}^{(i)}]}{[\mathbf{p}_j^{(i)}!]} \\
 &= \sum_{1 \leq |\mathbf{m}| \leq |\alpha|} f_{\mathbf{m}} \sum_{\pi \in p(\alpha, \mathbf{m})} \prod_{j=1}^{|\alpha|} \frac{[g_{\mathbf{p}_j}]^{m_j}}{(m_j!)[\mathbf{p}_j!]^{|m_j|}} \\
 &= \sum_{1 \leq |\mathbf{m}| \leq |\alpha|} \frac{m! f_{\mathbf{m}}}{\mathbf{m}!} \sum_{\pi \in p(\alpha, \mathbf{m})} \prod_{j=1}^{|\alpha|} \frac{[g_{\mathbf{p}_j}]^{m_j}}{(m_j!)[\mathbf{p}_j!]^{|m_j|}} \\
 &= \sum_{1 \leq |\mathbf{m}| \leq |\alpha|} \frac{f_{\mathbf{m}}}{\mathbf{m}!} \sum_{\pi \in s^+(\alpha, \mathbf{m})} \prod_{i=1}^{\mu} \prod_{j=1}^{r_i} \frac{[g_{\mathbf{p}_j}^{(i)}]}{[\mathbf{p}_j^{(i)}!]} \\
 &= \sum_{\pi \in s^+(\alpha)} \frac{f_{\mathbf{m}}}{\mathbf{m}!} \prod_{i=1}^{\mu} \prod_{j=1}^{r_i} \frac{[g_{\mathbf{p}_j}^{(i)}]}{[\mathbf{p}_j^{(i)}!]}
 \end{aligned}$$

We used formula 3.3 in going from the first line to the second, and formula 3.8 in going from the third to the fourth.  $\square$

### 4. More Bernoulli and Euler number identities

Recall from Section 1 that if

$$h_B(x_1, \dots, x_\nu) = \frac{x_1 + \dots + x_\nu}{e^{x_1 + \dots + x_\nu} - 1} \quad \text{and} \quad h_E(x_1, \dots, x_\nu) = \operatorname{sech}(x_1 + \dots + x_\nu),$$

then the  $\alpha$ -th generalized Bernoulli and Euler numbers are  $B_\alpha = \alpha! T_\alpha(h_B; \mathbf{0})$  and  $E_\alpha = \alpha! T_\alpha(h_E; \mathbf{0})$  respectively. In [Vella 2008], the Bernoulli and Euler number identities are expressed in terms of Stirling numbers of the second kind. In this section, we will derive more Bernoulli and Euler number identities using the multivariable analog of these Stirling numbers.

Recall the *multivariable Stirling number of the second kind*,

$$S(\alpha, \mathbf{m}) = \sum_{p(\alpha, \mathbf{m})} \alpha! \prod_{j=1}^{|\alpha|} \frac{1}{\mathbf{m}_j! (p_j!)^{\mathbf{m}_j}} = \sum_{p(\alpha, \mathbf{m})} \frac{\alpha!}{\lambda(\pi)! \pi!}, \tag{4}$$

introduced on page 516 of [Constantine and Savits 1996]. Additionally, we define

$$p(\alpha, \mathbf{m}, \text{even}) = \{(\mathbf{m}_1, \dots, \mathbf{m}_s; p_1, \dots, p_s) \in p(\alpha, \mathbf{m}) : |\mathbf{p}_j| \text{ even},$$

$$\text{for all } j \in \{1, \dots, s\}\},$$

$$s^+(\alpha, \mathbf{m}, \text{even}) = \{(\mathbf{p}_1^{(1)}, \dots, \mathbf{p}_{r_1}^{(1)}; \dots; \mathbf{p}_1^{(\mu)}, \dots, \mathbf{p}_{r_\mu}^{(\mu)}) \in s^+(\alpha, \mathbf{m}) : |\mathbf{p}_j^{(i)}| \text{ even},$$

$$\text{for all } i \in \{1, \dots, \mu\}, \text{ for all } j \in \{1, \dots, r_\mu\}\}.$$

We analogously define  $p(\alpha, \mathbf{m}, \text{odd})$  and  $s^+(\alpha, \mathbf{m}, \text{odd})$ . Let

$$p(\alpha, \text{even}) = \bigcup_{1 \leq |\mathbf{m}| \leq |\alpha|} p(\alpha, \mathbf{m}, \text{even}),$$

$$s^+(\alpha, \text{even}) = \bigcup_{1 \leq |\mathbf{m}| \leq |\alpha|} s^+(\alpha, \mathbf{m}, \text{even}),$$

and similarly define  $p(\alpha, \text{odd})$  and  $s^+(\alpha, \text{odd})$ . We call the  $p_i$  appearing in elements of  $p(\alpha, \text{even})$  and  $p(\alpha, \mathbf{m}, \text{even})$  *even parts* of  $\alpha$ , and we define *odd parts* of  $\alpha$  in the same manner. Furthermore, let

$$S(\alpha, \mathbf{m}, \text{even}) = \sum_{p(\alpha, \mathbf{m}, \text{even})} \alpha! \prod_{j=1}^{|\alpha|} \frac{1}{\mathbf{m}_j! (p_j!)^{\mathbf{m}_j}} = \sum_{p(\alpha, \mathbf{m}, \text{even})} \frac{\alpha!}{\lambda(\pi)! \pi!}, \tag{5}$$

and similarly define  $S(\alpha, \mathbf{m}, \text{odd})$ .

Our next theorem gives more explicit identities for calculating Bernoulli numbers.

**Theorem 4.1.** *If  $B_\alpha$  is the  $|\alpha|$ -th Bernoulli number, then*

$$(a) \quad B_\alpha = \sum_{\pi \in p(\alpha)} \frac{(-1)^m}{1+m} \binom{m}{\lambda(\pi)} \binom{\alpha}{\pi} = \sum_{\pi \in s^+(\alpha)} \frac{(-1)^m}{1+m} \binom{\alpha}{\pi},$$

$$(b) B_{\alpha} = \sum_{1 \leq m \leq |\alpha|} \frac{(-1)^m m!}{1+m} S(\alpha, m).$$

*Proof.* Let  $g(x_1, \dots, x_\nu) = e^{x_1 + \dots + x_\nu} - 1$  and  $f(y) = \ln(1+y)/y$ . Let  $\mathbf{x}^0 = \mathbf{0} \in \mathbb{R}^\nu$ . Then  $T_{\mathbf{p}_j}(g; \mathbf{0}) = 1/\mathbf{p}_j!$  if  $\mathbf{p}_j > \mathbf{0}$ , while  $T_m(f; \mathbf{y}^0) = T_m(f; \mathbf{0}) = (-1)^m/(1+m)$ . By Corollary 3.2,

$$T_{\alpha}(h; \mathbf{0}) = \sum_{1 \leq m \leq |\alpha|} \frac{(-1)^m}{1+m} \sum_{\pi \in p(\alpha, m)} \binom{m}{\lambda(\pi)} \prod_{j=1}^s \left[ \frac{1}{\mathbf{p}_j!} \right]^{m_j}.$$

Since  $B_{\alpha} = \alpha! T_{\alpha}(f \circ g; \mathbf{0})$ , this yields part (a) because

$$\begin{aligned} B_{\alpha} &= \alpha! T_{\alpha}(h; \mathbf{0}) = \sum_{1 \leq m \leq |\alpha|} \frac{(-1)^m}{1+m} \sum_{\pi \in p(\alpha, m)} \binom{m}{\lambda(\pi)} \frac{\alpha!}{\pi!} \\ &= \sum_{\pi \in p(\alpha)} \frac{(-1)^m}{1+m} \binom{m}{\lambda(\pi)} \binom{\alpha}{\pi} = \sum_{\pi \in s^+(\alpha)} \frac{(-1)^m}{1+m} \binom{\alpha}{\pi} \end{aligned}$$

by Proposition 3.4. Part (b) follows from part (a) because

$$m! S(\alpha, m) = \sum_{\pi \in p(\alpha, m)} \binom{m}{\lambda(\pi)} \binom{\alpha}{\pi}$$

by (4). Collecting together partitions of a fixed total multiplicity yields:

$$B_{\alpha} = \sum_{1 \leq m \leq |\alpha|} \frac{(-1)^m m!}{1+m} S(\alpha, m). \quad \square$$

Our next theorem gives more explicit identities for calculating Euler numbers.

**Theorem 4.2.** *If  $E_{\alpha}$  is the  $|\alpha|$ -th Euler number, then*

$$(a) E_{\alpha} = \sum_{\pi \in p(\alpha, \text{even})} (-1)^m \binom{m}{\lambda(\pi)} \binom{\alpha}{\pi} = \sum_{\pi \in s^+(\alpha, \text{even})} (-1)^m \binom{\alpha}{\pi},$$

$$(b) E_{\alpha} = \sum_{1 \leq m \leq |\alpha|} (-1)^m m! S(\alpha, m, \text{even}).$$

*Proof.* Let  $g(x_1, \dots, x_\nu) = \cosh(x_1, \dots, x_\nu)$  and  $f(y) = 1/y$ . Let  $\mathbf{x}^0 = \mathbf{0} \in \mathbb{R}^\nu$ . Then  $T_{\mathbf{p}_j}(g; \mathbf{0}) = 1/\mathbf{p}_j!$  for even parts and  $T_{\mathbf{p}_j}(g; \mathbf{0}) = 0$  for odd parts, while  $T_m(f; \mathbf{y}^0) = T_m(f; 1) = (-1)^m$ . From Corollary 3.2, we have

$$T_{\alpha}(h; \mathbf{0}) = \sum_{1 \leq m \leq |\alpha|} (-1)^m \sum_{\pi \in p(\alpha, m)} \binom{m}{\lambda(\pi)} \prod_{j=1}^s [T_{\mathbf{p}_j}(g; \mathbf{0})]^{m_j},$$

but if any of the parts of  $\pi$  are odd, the product vanishes. Thus, the sum becomes over partitions of only even parts, and



$$T_{\alpha}(h; \mathbf{0}) = \sum_{1 \leq m \leq |\alpha|} (-1)^m \sum_{\pi \in p(\alpha, m, \text{even})} \binom{m}{\lambda(\pi)} \prod_{j=1}^s \left[ \frac{1}{p_j!} \right]^{m_j}.$$

Since  $E_{\alpha} = \alpha! T_{\alpha}(h; \mathbf{0})$ , this yields part (a) because

$$\begin{aligned} E_{\alpha} &= \alpha! T_{\alpha}(h; \mathbf{0}) = \sum_{1 \leq m \leq |\alpha|} (-1)^m \sum_{\pi \in p(\alpha, m, \text{even})} \binom{m}{\lambda(\pi)} \frac{\alpha!}{\pi!} \\ &= \sum_{\pi \in p(\alpha, \text{even})} (-1)^m \binom{m}{\lambda(\pi)} \binom{\alpha}{\pi} = \sum_{\pi \in s^+(\alpha, \text{even})} (-1)^m \binom{\alpha}{\pi} \end{aligned}$$

by Proposition 3.4. Part (b) follows from part (a) because

$$m! S(\alpha, m) = \sum_{\pi \in p(\alpha, m)} \binom{m}{\lambda(\pi)} \binom{\alpha}{\pi}$$

by (5). Collecting together partitions of a fixed total multiplicity yields

$$E_{\alpha} = \sum_{1 \leq m \leq |\alpha|} (-1)^m m! S(\alpha, m, \text{even}). \quad \square$$

**Theorem 4.3.** *If  $E_{\alpha}$  is the  $|\alpha|$ -th Euler number, then*

$$1 = \sum_{1 \leq m \leq |\alpha|} \frac{(-1)^r}{(2r)!} E_{2r} \sum_{\pi \in p(\alpha, 2r, \text{odd})} \binom{2r}{\lambda(\pi)} \binom{\alpha}{\pi} \prod_{j=1}^s [E_{p_j}]^{m_j}.$$

*Proof.* Let  $g(x_1, \dots, x_v) = 2 \tan^{-1}(e^{x_1 + \dots + x_v}) - \pi/2$  be the multivariable analogue of the gudermannian function and set  $f(y) = \sec y$ . Let  $\mathbf{x}^0 = \mathbf{0}$ . Notice that  $h(x_1, \dots, x_v) = \sec(g(x_1, \dots, x_v)) = \cosh(x_1 + \dots + x_v)$ . Then  $T_{\alpha}(h; \mathbf{x}^0) = T_{\alpha}(h; \mathbf{0}) = 1/\alpha!$  when  $|\alpha|$  is even and  $T_{\alpha}(h; \mathbf{0}) = 0$  otherwise, while

$$T_m(f; \mathbf{y}^0) = T_m(f; \mathbf{0}) = \frac{(-1)^{m/2}}{m!} E_m$$

when  $m$  is even and  $T_m(f; \mathbf{0}) = 0$  when  $m$  is odd. Letting  $m = 2r$ , we substitute

$$T_{2r}(f; \mathbf{0}) = \frac{(-1)^r}{(2r)!} E_{2r}$$

into (2) of Corollary 3.2 to yield

$$\frac{1}{\alpha!} = \sum_{1 \leq 2r \leq |\alpha|} \frac{(-1)^r}{(2r)!} E_{2r} \sum_{\pi \in p(\alpha, 2r)} \binom{2r}{\lambda(\pi)} \prod_{j=1}^s [T_{p_j}(g; \mathbf{x}^0)]^{m_j}.$$

From the basic properties of the gudermannian function,

$$\begin{aligned} g(x_1, \dots, x_\nu) &= \int_0^{x_i} \operatorname{sech}(x_1 + \dots + x_\nu) dx_i \\ &= \sum_{j_1, \dots, j_\nu=0}^{\infty} \frac{E_{(j_1, \dots, j_\nu)}}{j_1! \dots j_\nu!} \int_0^{x_i} x_1^{j_1} \dots x_\nu^{j_\nu} dx_i \\ &= \sum_{j_1, \dots, j_\nu=0}^{\infty} \frac{E_{(j_1, \dots, j_\nu)}}{j_1! \dots (j_i + 1)! \dots j_\nu!} x_1^{j_1} \dots x_i^{j_i+1} \dots x_\nu^{j_\nu}. \end{aligned}$$

Thus,

$$T_{(j_1, \dots, j_i+1, \dots, j_\nu)}(g; \mathbf{x}^0) = T_{(j_1, \dots, j_i+1, \dots, j_\nu)}(g; \mathbf{0}) = \frac{E_{(j_1, \dots, j_\nu)}}{j_1! \dots (j_i + 1)! \dots j_\nu!}.$$

It follows that  $T_{(j_1, \dots, j_i+1, \dots, j_\nu)}(g; \mathbf{x}^0) = 0$  unless  $|(j_1, \dots, j_i + 1, \dots, j_\nu)|$  is odd because formula (a) of Theorem 4.2 implies that either  $E_{(j_1, \dots, j_\nu)} = 0$  or it is possible to write  $(j_1, \dots, j_\nu)$  as the sum of only even parts. It follows that

$$1 = \sum_{1 \leq m \leq |\alpha|} \frac{(-1)^r}{(2r)!} E_{2r} \sum_{\pi \in p(\alpha, 2r, \text{odd})} \binom{2r}{\lambda(\pi)} \binom{\alpha}{\pi} \prod_{j=1}^s [E_{p_j}]^{m_j}. \quad \square$$

### Acknowledgements

The author thanks David Vella for visiting Siena College to speak about the results of his work [2008], which inspired the current paper. The author also thanks Jon P. Bannon of Siena College for serving as a mentor to this research project and thanks Siena College for funding this project.

### References

- [Constantine and Savits 1996] G. M. Constantine and T. H. Savits, “A multivariate Faà di Bruno formula with applications”, *Trans. Amer. Math. Soc.* **348**:2 (1996), 503–520. MR 96g:05008 Zbl 0846.05003
- [Di Nardo and Oliva 2012] E. Di Nardo and I. Oliva, “Multivariate Bernoulli and Euler polynomials via Lévy processes”, *Appl. Math. Lett.* **25**:9 (2012), 1179–1184. MR 2930742 Zbl 1250.65015
- [Vella 2008] D. C. Vella, “Explicit formulas for Bernoulli and Euler numbers”, *Integers* **8** (2008), A01, 7. MR 2008j:11015 Zbl 1195.11033

Received: 2013-06-03      Revised: 2013-08-04      Accepted: 2013-09-24

fm20roma@siena.edu

*Department of Mathematics, Siena College,  
Loudonville, NY 12211, United States*

# Crossings of complex line segments

Samuli Leppänen

(Communicated by Kenneth S. Berenhaut)

The crossing lemma holds in  $\mathbb{R}^2$  because a real line separates the plane into two disjoint regions. In  $\mathbb{C}^2$  removing a complex line keeps the remaining point-set connected. We investigate the crossing structure of affine line segment-like objects in  $\mathbb{C}^2$  by defining two notions of line segments between two points and give computational results on combinatorics of crossings of line segments induced by a set of points. One way we define the line segments motivates a related problem in  $\mathbb{R}^3$ , which we introduce and solve.

## 1. Introduction

A graph is planar if it can be drawn on the plane such that none of its edges cross. For any graph  $G$ , we define the crossing number  $\text{cr}(G)$  to be the smallest possible number of edge crossings over all the planar drawings of  $G$ . In this paper, we will study and present some computational results in the two-dimensional complex plane motivated by the crossing number inequality. The crossing number inequality is a well-known tool in discrete geometry as it gives a lower bound for the crossing number of a graph [Ajtai et al. 1982]:

**Theorem 1.1** (crossing number inequality). *If an undirected graph with  $n$  vertices and  $m$  edges satisfies  $m > 4n$ , then we have  $\text{cr}(G) \geq m^3/64n^2$ .*

One of the applications of the inequality is a short proof [Székely 1997] of the Szemerédi–Trotter theorem [1983]:

**Theorem 1.2** (Szemerédi–Trotter theorem). *Given  $n$  points and  $m$  lines in the plane, the number of point-line pairs such that the point lies on the line is*

$$O(n^{2/3}m^{2/3} + n + m).$$

Theorem 1.2 generalizes to the two-dimensional complex plane [Tóth 2003] with lines of complex variable and points in the two-dimensional complex plane, and in a slightly weaker form to spaces of higher dimension [Solymosi and Tao 2012].

---

*MSC2010:* primary 51M05, 51M30, 52C35; secondary 51M04.

*Keywords:* discrete geometry, crossing inequality.

The main motivation of our work is the question of whether a suitable generalization of the crossing number inequality could yield a simple proof for the complex generalization of the Szemerédi–Trotter theorem in similar vein as in the real counterpart. The answer to this question is still out of reach and very little is known. One significant difficulty in understanding the crossing number of a graph in  $\mathbb{C}^2$  is that interpreting an edge in such a graph as a line segment is not as straightforward as in  $\mathbb{R}^2$ . One natural way to attempt to understand crossings of graphs in  $\mathbb{C}^2$  is to look for complete graphs without crossings. In  $\mathbb{R}^2$  it is well known that the complete graph with five or more vertices always has at least one crossing. Analogously, given a set of five or more points in  $\mathbb{R}^2$ , if we connect all the points with line segments, at least two of the line segments will cross. It is not clear to what extent the same is true in  $\mathbb{C}^2$ , and this will be the main focus of our study. In Section 2, we will present two ways to define a complex line segment and devise an algorithm that looks for configurations of  $n$  points such that the corresponding complete graph has no crossings. We will discuss the results and based on them give two conjectures regarding arrangements of points in  $\mathbb{C}^2$  and crossings of the line segments between them. In Section 3, we introduce and present a solution to a problem in  $\mathbb{R}^3$  motivated by our earlier discussion.

## 2. Line segments in $\mathbb{C}^2$

The two-dimensional complex plane is the set of points

$$\mathbb{C}^2 = \{(z_1, z_2) : z_1, z_2 \in \mathbb{C}\},$$

and a complex line determined by the constants  $a, b \in \mathbb{C}$  is the subset

$$\{(u, v) \in \mathbb{C}^2 : v = au + b\}.$$

The two-dimensional complex plane can be considered as a four-dimensional real Euclidean space with complex lines being two-dimensional affine subspaces. Since lines in  $\mathbb{C}^2$  are two-dimensional, it is not obvious how to define a line segment between two points  $z_1, z_2 \in \mathbb{C}^2$ . In general, we want a line segment to be a region enclosed by a simply connected curve on the complex line that contains the points  $z_1, z_2$ . For simplicity, we focus on two particular types of line segments: one given by the closed disk that has  $z_1$  and  $z_2$  as its antipodal points and another that is the union of the two closed disks centered at  $z_1$  and  $z_2$ , both having radius  $\|z_1 - z_2\|$ .

Before making these notions precise, let us briefly discuss the problem we will study: any arrangement of five points in  $\mathbb{R}^2$  is such that if we draw the line segments between all the points, then at least two of the line segments cross<sup>1</sup>. The same is

---

<sup>1</sup>By crossing of line segments we mean an intersection of two line segments that is not an endpoint of either line segment.

not true for every configuration of four points. This is equivalent to saying that the smallest complete graph with nonzero crossing number is the one with five vertices,  $K_5$ . We are interested in studying to what extent this is true for complex line segments in  $\mathbb{C}^2$ , or in particular, what is the number of points such that the induced line segments necessarily contain at least one crossing? We will present a computational algorithm that looks for configurations of points with no crossings for a given number of points. Using the algorithm, we can look for a lower bound for the number of points such that the induced graph does not have a crossing.

Let us denote the set of points in  $\mathbb{C}^2$  by

$$\begin{aligned} P &= \{z_1, z_2, \dots, z_n\} \\ &= \{(u_1, v_1), (u_2, v_2), \dots, (u_n, v_n)\}, \quad u_i, v_i \in \mathbb{C}, \end{aligned}$$

and a line containing the points  $z_i, z_j$  by

$$L_{ij} = \{(u, au + b) \in \mathbb{C}^2 : a, b \in \mathbb{C} \text{ s.t. } au_k + b = v_k, k = i, j\}.$$

We can now introduce the two notions of line segments.

**Definition.** Call the set

$$S_I(z_1, z_2) = \left\{ z \in L_{12} : \left\| z - \frac{z_1 + z_2}{2} \right\| \leq \left\| \frac{z_1 - z_2}{2} \right\| \right\}$$

a textline segment of type I.

**Definition.** Call the set

$$S_{II}(z_1, z_2) = \{z \in L_{12} : \|z - z_1\| \leq \|z_1 - z_2\| \text{ or } \|z - z_2\| \leq \|z_1 - z_2\|\}$$

a line segment of type II.

If the type of the line segment is irrelevant, we will just write  $S(z_1, z_2)$ . We say that the line segments  $S(z_i, z_j)$  and  $S(z_k, z_l)$  (where no two points are equal) have a crossing if and only if

$$S(z_i, z_j) \cap S(z_k, z_l) = L_{ij} \cap L_{kl} \neq \emptyset.$$

**Computational setup.** We observe that if two line segments do not cross, then the intersection point of the lines defined by the points lies outside of at least one of the line segments. This motivates us to look for configurations of points where the intersection point of any two lines is in some sense close to the boundary of the curve defining the line segment.

Let  $z_i, z_j, z_k, z_l$  be distinct points of the set  $P$ . Denote by  $z = L_{ij} \cap L_{kl}$  the intersection of one of the pairs of lines induced by the points. For an intersection

of line segments of type I, set

$$r_{ij}^I = \frac{\|z - \frac{z_i + z_j}{2}\|}{\frac{1}{2}\|z_i - z_j\|}$$

to measure the relative distance of the intersection point from the center of the circle defining the line segment. For the lines  $L_{ij}$  and  $L_{kl}$ , define

$$\rho_{ij,kl}^I = \max\{r_{ij}^I, r_{kl}^I\}.$$

For each pair of lines,  $\rho_{ij,kl}^I$  picks the one for which the intersection point of the lines is relatively further from the center of the circle defining the line segment. Finally, set

$$\rho^I = \min_{z_i, z_j, z_k, z_l \in P} \{\rho_{ij,kl}^I, \rho_{ik,jl}^I, \rho_{il,jk}^I\},$$

where all the points  $z_i, z_j, z_k, z_l$  are distinct. Similarly, for an intersection of line segments of type II, set

$$r_{ij}^{II} = \min \left\{ \frac{\|z - z_i\|}{\|z_i - z_j\|}, \frac{\|z - z_j\|}{\|z_i - z_j\|} \right\},$$

and define the quantities  $\rho_{ij,kl}^{II}$  and  $\rho^{II}$  in the same way we did for the line segment of type I. In what follows, we will just write  $\rho$  instead of  $\rho^I$  or  $\rho^{II}$  when it does not matter which type of line segment is in question. Furthermore, notice that  $\rho$  is a function of the set of points  $P$ , but to simplify notation we will leave it unwritten.

Evidently if  $\rho > 1$ , none of the line segments defined by the points in the configuration have a crossing. We will use a randomized algorithm to search for configurations with  $\rho$  close to 1 in hope of either finding a configuration that contains no crossing of the induced line segments or a configuration that is extremal in the sense that  $\rho \approx 1$ .

The way our algorithm works is as follows: Initially start with a random configuration  $P_0 = \{z_1, \dots, z_n\}$ . On iteration  $k$ , choose an index  $j \in \{1, \dots, n\}$  randomly using a uniform distribution and set  $\hat{z}_j = z_j + \epsilon$ , where  $\epsilon \in \mathbb{C}^2$  is some uniformly distributed random variable with 0 mean and small variance. If the  $\rho$  computed for the new configuration is larger than the  $\rho$  of the configuration from the previous iteration, replace  $z_j$  by  $\hat{z}_j$  in the configuration, otherwise do nothing.

In order to justify the algorithm, let us make the following remarks: The results of the described algorithm provide us with lower bounds for the number of points whose induced complete graph does not necessarily have a crossing. The algorithm makes small local perturbations to maximize the quantity  $\rho$ , but it is not clear whether or not there are several local optima that differ from a global optimum. Therefore, the cases where the algorithm fails to find a noncrossing configuration

are inconclusive. However, when applied to  $\mathbb{R}^2$ , the algorithm found noncrossing configurations for four points but not for five, agreeing with known results.

**Results.** Our computational experiments motivate the following remark and two conjectures:

**Remark.** There is a configuration of seven points in  $\mathbb{C}^2$  such that none of the line segments of type I between any pairs of points have a crossing.

One such configuration, with  $\rho^I \approx 1.1047$ , is

$$\begin{aligned} z_1 &= (0.4358 - 0.3796i, 0.5726 + 0.3896i), \\ z_2 &= (-0.3382 + 0.0719i, -0.1316 + 0.3220i), \\ z_3 &= (0.6391 + 0.0141i, 0.8889 - 0.3292i), \\ z_4 &= (0.6302 - 0.5513i, 0.2813 - 0.8285i), \\ z_5 &= (0.9731 - 1.3291i, 2.3615 + 0.4571i), \\ z_6 &= (1.7105 - 0.7780i, -1.4009 - 0.8982i), \\ z_7 &= (0.0099 - 0.9417i, 1.3350 - 0.9040i). \end{aligned}$$

We were not able to produce a configuration of eight points such that  $\rho^I \geq 1$ . We observed that when executing the search algorithm with 20000 iterations ten times,  $\rho^I$  was found to lie between 0.978347 and 0.999998. Hence we state the following conjecture:

**Conjecture.** *Every configuration of eight points in  $\mathbb{C}^2$  has four points such that the line segments of type I induced by the points have an intersection. In particular, there exists a configuration of eight points such that  $\rho^I = 1$ .*

For line segments of type II, we were not able to produce a configuration of four points such that  $\rho^{II} > 1$  after executing the search algorithm with 20000 iterations ten times. We noticed that there exists a configuration such that  $\rho^{II} = 1$ ; for example, consider the points

$$\begin{aligned} z_1 &= (0, 0), \\ z_2 &= (1, 0), \\ z_3 &= \left(\frac{1}{2} + \frac{\sqrt{3}}{2}i, 0\right), \\ z_4 &= (u, v), \quad \text{where } u, v \in \mathbb{C}, v \neq 0. \end{aligned}$$

It is not difficult to see that this configuration has the claimed property, as  $z_1, z_2$  and  $z_3$  all lie on the same complex line and have equal distance from each other. Thus the following conjecture is motivated:

**Conjecture.** *Every configuration of four points in  $\mathbb{C}^2$  is such that at least two of the line segments of type II induced by the points have an intersection.*

### 3. A related problem in $\mathbb{R}^3$

Line segments of type I define a disk with two given points as antipodal points. In the above treatment, we were interested in configurations of points in  $\mathbb{C}^2$  such that the line segments between the points do not intersect. This motivates a similar question in  $\mathbb{R}^3$ , which we will introduce and produce a solution for.

Consider a set of  $n$  points  $P = \{p_1, p_2, \dots, p_n\} \subset \mathbb{R}^3$ . For each pair of points  $p_i, p_j$ , denote by  $T_{ij}$  some plane containing both points and by  $D_{ij}$  the closed disk lying on  $T_{ij}$  with antipodal points  $p_i, p_j$ . In other words,

$$D_{ij} = \left\{ x \in \mathbb{R}^3 : x \in T_{ij}, \left\| x - \frac{p_i - p_j}{2} \right\| \leq \left\| \frac{p_i - p_j}{2} \right\| \right\}.$$

We will call  $\mathcal{D} = \{D_{ij} : i < j, i, j = 1, \dots, n\}$  a *disk system induced by  $P$* . For a pair of such disks,  $D_{ij}, D_{kl} \in \mathcal{D}$ , we say that the disks *intersect properly* if  $D_{ij} \cap D_{kl} \not\subseteq P$ . Fixing the set  $P$  does not trivially determine if there is a pair of disks that intersect properly in  $\mathcal{D}$  since there is some freedom in choosing each of the planes  $T_{ij}$  (i.e., the rotation of the disk  $D_{ij}$  around the line passing through  $p_i$  and  $p_j$ ). We are now interested in determining the conditions for the set  $P$  such that none of the pairs of disks intersect properly. In what follows, we prove the following result:

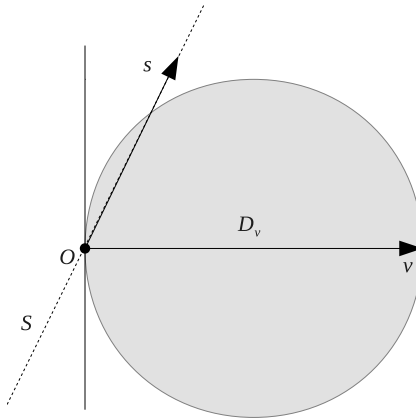
**Theorem 3.1.** *The maximal size of the set  $P$  such that the induced disks do not intersect properly is four. In such a configuration all the points lie on a plane  $T$ , and three of the points form a triangle with one point in its interior. All the disks intersect  $T$  perpendicularly.*

**Remark.** Notice the differences between line segments of type I we defined in Section 2 and the disks considered here: the line segments of type I reside in four-dimensional space and their rotation along the axis given by the two points is fixed. In addition, when considering the proper intersections of the disks  $D_{ij}$  and  $D_{kl}$  here, we do not require that  $i, j, k, l$  are all different.

**Proofs.** We will first characterize proper intersections of two disks sharing a common point. Then using this characterization, we show that for three points, there is only one way of choosing the rotations of the disks such that no two intersect properly, which quickly implies Theorem 3.1.

*Two disks.* To keep notation simple, let  $v, w \in \mathbb{R}^3$  be two nonparallel vectors. Let  $T_v, T_w$  be two planes such that  $T_v$  is spanned by  $v$  and some (still unspecified) vector, and  $T_w$  is similarly spanned by  $w$  and some other vector. Denote by  $D_v$  the disk lying in  $T_v$  such that the antipodal points of  $D_v$  are the origin and  $v$ , and by  $D_w$  the disk lying in  $T_w$  with the origin and  $w$  as antipodal points.





**Figure 1.** The disk  $D_v$ , line  $S$  and its spanning vector  $s$  on the  $T_v$ -plane.

Since  $T_v$  and  $T_w$  both contain the origin, their intersection is always nonempty. Let  $S = T_v \cap T_w$  be the line given by the intersection of the two planes and  $s$  a vector such that  $S = \text{span } s$ . Ignoring the trivial case of  $\text{span } v = \text{span } s$  or  $\text{span } w = \text{span } s$ , we have that  $T_v = \text{span}(v, s)$  and  $T_w = \text{span}(w, s)$ . Therefore, the disks  $D_v$ ,  $D_w$  and thus their intersection is determined by the three vectors  $v$ ,  $w$  and  $s$ .

The line  $S$  is given by the intersection of the planes  $T_v$  and  $T_w$ , but what does it tell us about the intersection of the disks? First, let us see how things look on the  $T_v$ -plane (see Figure 1). If  $s$  is perpendicular to  $v$ , then clearly the disk  $D_v$  does not intersect the plane  $T_w$  outside of the origin and hence cannot intersect  $D_w$  properly. Otherwise it is clear that there exists some real  $\alpha \neq 0$  such that  $\alpha s \in D_v$ , i.e.,  $S$  intersects  $D_v$  outside the origin.

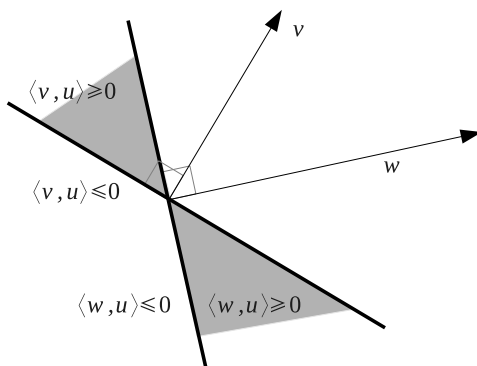
The same conclusion naturally holds for the disk  $D_w$ . Let us use this observation to prove the following lemma:

**Lemma 3.2.** *The disks  $D_v$  and  $D_w$  intersect properly if and only if*

$$\langle v, s \rangle \langle w, s \rangle > 0.$$

*Proof.* If  $D_v$  and  $D_w$  intersect properly, there is some nonzero  $\alpha \in \mathbb{R}$  such that  $\alpha s \in D_v \cap D_w$  since the intersection  $S \cap D_v \cap D_w$  is not just the origin. Then, from the way we have defined the disks  $D_v$ ,  $D_w$  to lie on the planes  $T_v$ ,  $T_w$  (see Figure 1), it follows that the projection of  $\alpha s$  to the vector  $v$  has the same direction as  $v$ , and the projection of  $\alpha s$  to  $w$  has the same direction as  $w$ . In other words,  $\langle v, \alpha s \rangle > 0$  and  $\langle w, \alpha s \rangle > 0$ . Multiplying these two inequalities together yields  $\alpha^2 \langle v, s \rangle \langle w, s \rangle > 0$ .

On the other hand, if  $\langle v, s \rangle \langle w, s \rangle > 0$ , then either  $\langle v, s \rangle$  and  $\langle w, s \rangle$  are both strictly positive or negative. Assume they are both positive. This means that for an arbitrarily small  $\alpha > 0$ , we must have  $\alpha s \in D_v$  and  $\alpha s \in D_w$ , i.e.,  $\alpha s \in D_v \cap D_w$ ,



**Figure 2.** The region of all the vectors  $u$  satisfying  $\langle v, u \rangle \langle w, u \rangle \leq 0$  on the plane spanned by  $v, w$  (shaded).

so the intersection of the disks contains points other than the origin. If both of the inner products are negative, the same conclusion holds for  $-\alpha$ .  $\square$

To see one useful interpretation of the above lemma, let us consider the orthogonal projection  $s'$  of  $s$  to the plane  $T = \text{span}(v, w)$ . First, note that  $\langle v, s \rangle = \langle v, s' \rangle$  and  $\langle w, s \rangle = \langle w, s' \rangle$ , so

$$\langle v, s \rangle \langle w, s \rangle = \langle v, s' \rangle \langle w, s' \rangle.$$

Therefore, the set of vectors  $s$  such that the disks  $D_v, D_w$  do not intersect properly, i.e.,  $\langle v, s \rangle \langle w, s \rangle \leq 0$ , is characterized by the cone  $C$  in  $T$  (see Figure 2), where

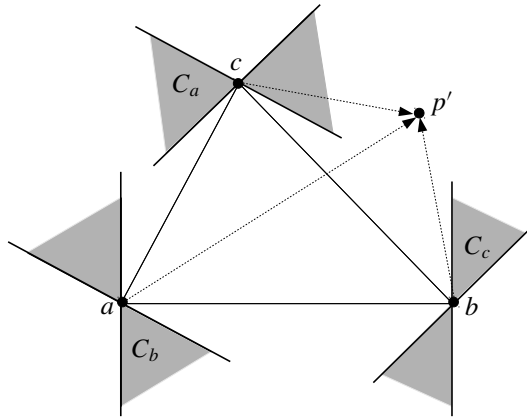
$$C = \{u \in T : \langle u, v \rangle \leq 0 \text{ and } \langle u, w \rangle \geq 0 \text{ or } \langle u, v \rangle \geq 0 \text{ and } \langle u, w \rangle \leq 0\}.$$

*Three disks.* We will now look at the implications of Lemma 3.2 to configurations of three disks. First we will show a fact from plane geometry concerning triangles and cones. Let  $a, b, c$  be noncollinear points on the plane and  $abc$  the corresponding triangle. To each vertex of the triangle we can associate a cone, as in Figure 2. Let the cones associated with the points  $a, b$  and  $c$  be called  $C_a, C_b$  and  $C_c$  (see Figure 3).

**Lemma 3.3.** *The intersection of the cones is empty, that is,  $C_a \cap C_b \cap C_c = \emptyset$ .*

*Proof.* Assume that the angle  $\alpha = \angle bac$  corresponding to the point  $a$  is the largest angle of the triangle. Since the opening angle of the cone is the same as the corresponding angle in the triangle, the opening angle of  $C_a$  is greater than the opening angles of  $C_b$  and  $C_c$ . Denote by  $l$  the line passing through the point  $a$  such that  $l$  halves the angle  $\alpha$ . Then  $l$  divides the plane into two parts, one containing the point  $b$  and once containing the point  $c$ ; denote these half-planes by  $H_b$  and  $H_c$ .

Since the opening angle of  $C_a$  is greater than the opening angle of  $C_b$  and  $C_c$ , and the opening angles of the cones are equal to the corresponding angles in the triangle, we have that the opening angles of  $C_b$  and  $C_c$  are strictly less than  $\pi/2$ . Thus  $C_b$



**Figure 3.** The triangle  $abc$ , the projection  $p'$  of  $p$  and the cones  $C_a$ ,  $C_b$  and  $C_c$ .

or  $C_c$  cannot contain any points in the triangle and so  $a \notin C_b \cup C_c$ . Therefore, the intersection  $C_a \cap C_b$  is entirely contained in  $H_b \setminus l$  and the intersection  $C_a \cap C_c$  in  $H_c \setminus l$ .  $\square$

Now we can show that there is only one way three points can induce a disk system without proper intersections. The points  $a, b, c$  lie on a plane  $T$  and determine a triangle  $abc$ . Each vertex of the triangle is a touching point of two disks, and each side of the triangle is the rotation axis of one disk. The rotation of a disk determines a plane containing the corresponding side of the triangle. If none of the three planes are equal, there are exactly two different cases for their intersection: either all three planes intersect in one point, or they are all perpendicular to the plane  $T$  and thus do not have a common intersection point.

We will show now that if the three planes have a mutual intersection point, then at least two of the disks will intersect properly. So assume there is a point  $p$  where the three planes intersect, and consider the orthogonal projection  $p'$  of  $p$  onto the plane  $T$  containing the points  $a, b, c$  (see Figure 3). As we saw earlier, if two disks touching in one vertex of the triangle do not intersect properly, then the line segment from the vertex to  $p'$  lies in the cone associated with the vertex. So to require that none of the pairs of disks intersect is the same as requiring that  $p' \in C_a \cap C_b \cap C_c$ , which by Lemma 3.3 is not possible.

We have justified the following:

**Lemma 3.4.** *The only disk system induced by three points such that no two disks intersect properly is the one where all the disks perpendicularly intersect the plane containing the points.*

Theorem 3.1 follows now without much effort. First, assume there is a configuration of four points  $p_1, \dots, p_4$  such that no two disks intersect properly and all the

points do not lie on the same plane. Then by Lemma 3.4, the disks induced by  $p_1$ ,  $p_2$  and  $p_3$  must all perpendicularly intersect the plane  $T$  containing  $p_1$ ,  $p_2$  and  $p_3$ . But the points  $p_1$ ,  $p_2$  and  $p_4$  lie on a plane  $T' \neq T$ , and the induced disks have to intersect  $T'$  perpendicularly. Therefore  $D_{12}$  intersects  $T$  and  $T'$  perpendicularly, which leaves no option other than  $T = T'$ , which contradicts our assumption.

Hence, for any number of points, we have to have that the points lie on a plane in order to not have properly intersecting disks in the induced disk system. The points and the disks give rise to a complete graph on the plane, as we can think of the points as vertices and the rotation axes as edges of the graph. Clearly the disks intersect properly if the graph has crossing edges. Any complete graph with five or more vertices has an edge crossing, which concludes the proof of Theorem 3.1.

### Acknowledgements

I wish to express my gratitude to my advisor József Solymosi for his support and ideas for this project.

### References

- [Ajtai et al. 1982] M. Ajtai, V. Chvátal, M. M. Newborn, and E. Szemerédi, “Crossing-free subgraphs”, pp. 9–12 in *Theory and practice of combinatorics*, edited by A. Rosa et al., North-Holland Math. Stud. **60**, North-Holland, Amsterdam, 1982. MR 86k:05059 Zbl 0502.05021
- [Solymosi and Tao 2012] J. Solymosi and T. Tao, “An incidence theorem in higher dimensions”, *Discrete Comput. Geom.* **48**:2 (2012), 255–280. MR 2946447 Zbl 1253.51004
- [Székely 1997] L. A. Székely, “Crossing numbers and hard Erdős problems in discrete geometry”, *Combin. Probab. Comput.* **6**:3 (1997), 353–358. MR 98h:52030 Zbl 0882.52007
- [Szemerédi and Trotter 1983] E. Szemerédi and W. T. Trotter, Jr., “Extremal problems in discrete geometry”, *Combinatorica* **3**:3-4 (1983), 381–392. MR 85j:52014 Zbl 0541.05012
- [Tóth 2003] C. Tóth, “The Szemerédi–Trotter theorem in the complex plane”, preprint, 2003. arXiv math/0305283v5

Received: 2013-07-16      Revised: 2014-02-22      Accepted: 2014-02-23

psleppanen@gmail.com

*Department of Mathematics, University of British Columbia,  
Vancouver BC V6T 1Z2, Canada*

# On the $\varepsilon$ -ascent chromatic index of complete graphs

Jean A. Breytenbach and C. M. (Kieka) Mynhardt

(Communicated by Jerrold Griggs)

An edge ordering of a graph  $G = (V, E)$  is an injection  $f : E \rightarrow \mathbb{Z}^+$ , where  $\mathbb{Z}^+$  is the set of positive integers. A path in  $G$  for which the edge ordering  $f$  increases along its edge sequence is called an  $f$ -ascent; an  $f$ -ascent is maximal if it is not contained in a longer  $f$ -ascent. The depression  $\varepsilon(G)$  of  $G$  is the smallest integer  $k$  such that any edge ordering  $f$  has a maximal  $f$ -ascent of length at most  $k$ . Applying the concept of ascents to edge colourings rather than edge orderings, we consider the problem of determining the minimum number  $\chi_\varepsilon(K_n)$  of colours required to edge colour  $K_n$ ,  $n \geq 4$ , such that the length of a shortest maximal ascent is equal to  $\varepsilon(K_n) = 3$ . We obtain new upper and lower bounds for  $\chi_\varepsilon(K_n)$ , which enable us to determine  $\chi_\varepsilon(K_n)$  exactly for  $n = 7$  and  $n \equiv 2 \pmod{4}$  and to bound  $\chi_\varepsilon(K_{4m})$  by  $4m \leq \chi_\varepsilon(K_{4m}) \leq 4m + 1$ .

## 1. Introduction

Following [Schurch 2013a; 2013b], we consider the following question:

**Question 1.** For  $n \geq 4$ , what is the smallest integer  $r(n)$  for which there exists a proper edge colouring of  $K_n$  in colours  $1, \dots, r(n)$  such that a shortest maximal path of increasing edge labels has length three?

Schurch showed that  $r(n) \leq 2n - 3$  for all  $n \geq 4$ . This bound enabled him to determine  $r(n)$  for  $n \in \{4, 5\}$  and to show that  $7 \leq r(6) \leq 8$ . In Section 2 we give a lower bound for  $r(n)$  and in Section 3 we improve the general upper bound to

$$r(n) \leq \left\lfloor \frac{3n-3}{2} \right\rfloor.$$

We then improve this bound for even values of  $n$ . Consequently, we obtain  $r(7) = 9$ ,  $r(n) = n + 1$  if  $n \equiv 2 \pmod{4}$ , and  $n \leq r(n) \leq n + 1$  if  $n \equiv 0 \pmod{4}$  and  $n \geq 8$ .

*MSC2010:* 05C15, 05C78, 05C38.

*Keywords:* edge ordering of a graph, increasing path, depression, edge colouring.

Breytenbach was a second year undergraduate student, enrolled for the degree BSc in Mathematical Sciences (stream Computer Science) at Stellenbosch University, while this paper was being prepared. The paper earned him extra credit for the Foundations of Abstract Mathematics I course. Mynhardt was supported by an NSERC discovery grant.

We begin with a short historical account of the background to this problem. An *edge ordering* of a finite, simple graph  $G$  is an injection  $f : E(G) \rightarrow \mathbb{Z}^+$ , where  $\mathbb{Z}^+$  is the set of positive integers. Denote the set of all edge orderings of  $G$  by  $\mathcal{F}(G)$ . A path  $v_1, \dots, v_k$  (where  $v_k \neq v_1$ ) in  $G$  such that  $f(v_1) < \dots < f(v_k)$  is called an *f-ascent*; an *f-ascent* is *maximal* if it is not contained in a longer *f-ascent*. The *height*  $H(f)$  of an edge ordering  $f$  is the length of a longest *f-ascent*, and the *flatness* of  $f$ , denoted by  $h(f)$ , is the length of a shortest maximal *f-ascent* of  $G$ .

Chvátal and Komlós [1971] posed the problem of determining

$$\alpha(K_n) = \min_{f \in \mathcal{F}(K_n)} \{H(f)\}$$

of the complete graph  $K_n$ . This is a difficult problem and  $\alpha(K_n)$  is known only for  $1 \leq n \leq 8$  (see [Burger et al. 2005; Chvátal and Komlós 1971]). The parameter  $\alpha(G)$  for complete and other finite graphs was also investigated in [Bialostocki and Roditty 1987; Burger et al. 2005; Calderbank et al. 1984; Graham and Kleitman 1973; Mynhardt et al. 2005; Roditty et al. 2001; Yuster 2001].

For an arbitrary finite graph  $G$ , Cockayne et al. [2006] considered the problem of determining  $\varepsilon(G) = \max_{f \in \mathcal{F}(G)} \{h(f)\}$ , that is, the maximum length, taken over all edge orderings  $f \in \mathcal{F}(G)$ , of a shortest maximal *f-ascent*. The parameter  $\varepsilon(G)$  is known as the *depression* of  $G$  and its computation is likewise a difficult problem. Another interpretation of the depression of  $G$  is that any edge ordering  $f$  of  $G$  has a maximal *f-ascent* of length at most  $\varepsilon(G)$ , and  $\varepsilon(G)$  is the smallest integer for which this statement is true. Graphs with depression two were characterized in [Cockayne et al. 2006], while trees with depression three were characterized in [Mynhardt 2008]. Graphs with no adjacent vertices of degree three or higher that have depression three were characterized in [Mynhardt and Schurch 2013]. Further work on depression can be found in [Cockayne and Mynhardt 2006; Gaber-Rosenblum and Roditty 2009; Schurch and Mynhardt 2014; 2014; Schurch 2013a; 2013b].

An edge ordering of  $G$  is also a *proper edge colouring* — a labelling of the edges of  $G$  such that adjacent edges have different labels. The minimum number of labels, also called *colours*, is called the *edge chromatic number* or the *chromatic index*  $\chi'(G)$ . It is well known (see [Chartrand et al. 2011, Section 10.2], for example) that  $\chi'(K_n) = n - 1$  if  $n$  is even and  $\chi'(K_n) = n$  if  $n$  is odd. A 1-*factor* of  $G$  is a 1-regular spanning subgraph of  $G$ , and  $G$  is 1-*factorable* if  $E(G)$  can be partitioned into 1-factors. If  $G$  is 1-factorable, then  $G$  is  $r$ -regular for some  $r$  and  $\chi'(G) = r$ . König's theorem (see [Chartrand et al. 2011, Theorem 10.15]) states that every  $r$ -regular bipartite graph is 1-factorable. In particular, the chromatic index of the complete bipartite graph  $K_{n,n}$  is given by  $\chi'(K_{n,n}) = n$ .

Noticing that the labels of some edges in an edge ordering of  $G$  may be unimportant when determining  $\varepsilon(G)$ , Schurch applied the concept of ascents to edge

colourings and called the minimum number of colours in a proper edge colouring  $c$  of  $G$  such that  $h(c) = \varepsilon(G)$  the  $\varepsilon$ -ascent chromatic index of  $G$ , denoted  $\chi_\varepsilon(G)$ . Unlike the case for general graphs, the depression of  $K_n$  is easy to determine:  $\varepsilon(K_1) = 0$ ,  $\varepsilon(K_2) = 1$ ,  $\varepsilon(K_3) = 2$  and  $\varepsilon(K_n) = 3$  for all  $n \geq 4$  (see [Cockayne et al. 2006]); that is, there does not exist an edge ordering or an edge colouring of  $K_n$  such that a shortest maximal ascent has length four or more. Note that  $\chi_\varepsilon(K_1) = 0$ ,  $\chi_\varepsilon(K_2) = 1$ ,  $\chi_\varepsilon(K_3) = 3$ , and determining  $\chi_\varepsilon(K_n)$  for  $n \geq 4$  is equivalent to finding the smallest integer  $r(n)$  such that there exists a proper edge colouring  $c$  of  $K_n$  in colours  $1, \dots, r(n)$  with  $h(c) = 3$ , as formulated in Question 1.

**2. Lower bound for the  $\varepsilon$ -ascent chromatic index of  $K_n$**

We begin with a simple lower bound for  $\chi_\varepsilon(K_n)$ , which slightly improves the bound in [Schurch 2013b, Proposition 8] in the special case where  $G = K_n$ .

**Theorem 1.** *If  $n \geq 4$ , then*

$$\chi_\varepsilon(K_n) \geq \begin{cases} n & \text{if } n \equiv 0 \pmod{4}, \\ n + 1 & \text{if } n \equiv 1, 2 \pmod{4}, \\ n + 2 & \text{if } n \equiv 3 \pmod{4}. \end{cases}$$

*Proof.* Let  $c$  be a proper edge colouring of  $K_n$  in colours  $1, \dots, r$  such that  $h(c) = 3$ . Such a colouring exists because  $\varepsilon(K_n) = 3$  if  $n \geq 4$ . For  $i = 1, \dots, r$ , define

$$E_i = \{e \in E(K_n) : c(e) = i\}.$$

Then  $|E_i| \leq \lfloor n/2 \rfloor$  for each  $i$ . Also, no vertex  $v$  is incident with an edge  $e \in E_1$  and an edge  $e' \in E_r$ , otherwise  $e, e'$  is a maximal  $c$ -ascent of length two, which contradicts  $h(c) = 3$ . Thus  $|E_1 \cup E_r| \leq \lfloor n/2 \rfloor$  and  $E_1 \cup E_r$  is an independent set of edges, that is,  $E_1 \cup E_r, E_2, \dots, E_{r-1}$  is also a proper edge colouring of  $K_n$ . Hence  $r \geq \chi'(K_n) + 1$ . In particular,

$$\chi_\varepsilon(K_n) \geq \begin{cases} n & \text{if } n \equiv 0 \pmod{4}, \\ n + 1 & \text{if } n \equiv 1 \pmod{4}. \end{cases}$$

Assume  $n \equiv 2 \pmod{4}$ ; say  $n = 4p + 2$ . Then  $K_n$  has  $(2p + 1)(4p + 1)$  edges. Suppose  $r = \chi'(K_n) + 1 = n$ . The upper bound

$$|E_1 \cup E_r|, |E_2|, \dots, |E_{r-1}| \leq \left\lfloor \frac{n}{2} \right\rfloor$$

implies that

$$|E_1 \cup E_r| = |E_2| = \dots = |E_{r-1}| = \left\lfloor \frac{n}{2} \right\rfloor = 2p + 1.$$

Since  $|E_1| + |E_r| = 2p + 1$ , an odd number,  $|E_1| \neq |E_r|$ . Without loss of generality say  $|E_1| = k$ , where  $k \leq p$ , and  $|E_r| = 2p + 1 - k$ . Suppose  $e \in E_2$  is not adjacent to

any edge in  $E_1$ . Since  $|E_1 \cup E_r| = 2p + 1 = \lfloor n/2 \rfloor$ ,  $e$  is adjacent to an edge  $e' \in E_r$ . But then  $e, e'$  is a maximal  $c$ -ascent of length two, which contradicts  $h(c) = 3$ . Therefore each edge in  $E_2$  is adjacent to an edge in  $E_1$ , and since  $c$  is a proper edge colouring,  $|E_2| \leq 2|E_1| = 2k \leq 2p < \lfloor n/2 \rfloor$ , a contradiction. Thus  $r \geq n + 1$  as required.

Assume  $n \equiv 3 \pmod{4}$ ; say  $n = 4p + 3$ . Then  $|E(K_n)| = (4p + 3)(2p + 1)$ . Suppose  $r = \chi'(K_n) + 1 = n + 1$ . As in the case  $n \equiv 2 \pmod{4}$ , we obtain that  $|E_1 \cup E_r| = |E_2| = \dots = |E_{r-1}| = \lfloor n/2 \rfloor = 2p + 1$  and that each edge in  $E_2$  is adjacent to an edge in  $E_1$ . There is one vertex  $v$  that is not incident with any edge in  $E_1 \cup E_r$ , but an edge in  $E_2$  incident with  $v$  also needs to be adjacent to an edge in  $E_1$ . We obtain a contradiction as above and the result follows.  $\square$

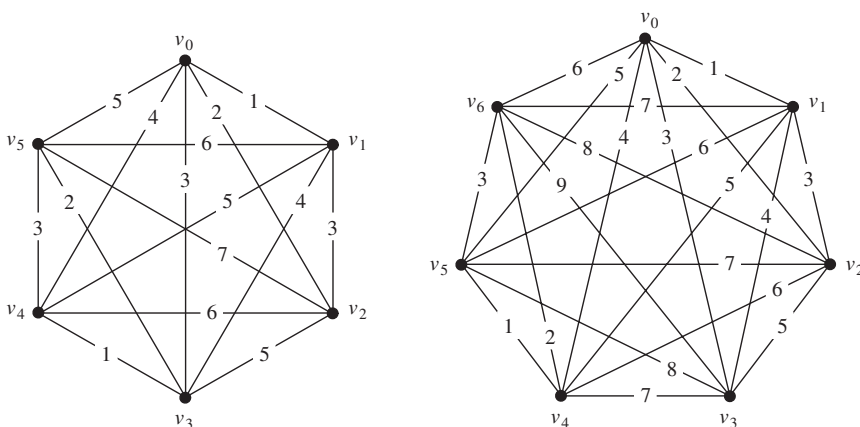
### 3. Upper bounds for the $\varepsilon$ -ascent chromatic index of $K_n$

In Section 3.1 we provide a new general upper bound for  $\chi_\varepsilon(K_n)$ . We improve this bound for even values of  $n$  in Sections 3.2 (the case  $n \equiv 0 \pmod{4}$ ) and 3.3 (the case  $n \equiv 2 \pmod{4}$ ).

**3.1. A general bound.** For  $n \geq 6$ , we now describe an edge colouring  $c$  of  $K_n$  in  $\lfloor (3n - 3)/2 \rfloor$  colours, as illustrated in Figure 1 for  $n \in \{6, 7\}$ , and prove in Theorem 3 that  $h(c) = 3$ . Let  $V(K_n) = \{v_0, \dots, v_{n-1}\}$  and  $p = \lceil n/2 \rceil$ .

- For  $i \in \{0, \dots, p - 1\}$  and  $j \in \{i + 1, \dots, n - 1\}$ , let  $c(v_i v_j) = i + j$ .
- For  $i \in \{p, \dots, n - 2\}$  and  $j \in \{i + 1, \dots, n - 1\}$ , let  $c(v_i v_j) = i + j - 2p$ .

**Lemma 2.** For all  $n \geq 6$ , the colouring  $c$  defines a proper edge colouring of  $K_n$  in  $\lfloor (3n - 3)/2 \rfloor$  colours.



**Figure 1.** Edge colourings of  $K_6$  and  $K_7$  with flatness three.



*Proof.* Suppose that  $c(v_i v_j) = c(v_i v_{j'})$  for some  $j < j'$ . After a brief reflection, we deduce that  $i + j = i + j' - 2p$ . But  $i + j \geq i$  and

$$i + j' - 2p \leq i + n - 1 - 2\lceil n/2 \rceil \leq i - 1,$$

hence  $c(v_i v_j) > c(v_i v_{j'})$ , contradicting our assumption.

Since the smallest colour is  $0 + 1 = 1$  and the largest colour is

$$p - 1 + n - 1 = \left\lceil \frac{n}{2} \right\rceil + n - 2 = \left\lfloor \frac{n-1}{2} \right\rfloor + n - 1 = \left\lfloor \frac{3n-3}{2} \right\rfloor,$$

the colouring  $c$  uses exactly  $\lfloor (3n - 3)/2 \rfloor$  colours. □

**Theorem 3.** *For all  $n \geq 6$ , the colouring  $c$  of  $K_n$  has flatness equal to three.*

*Proof.* To prove that  $h(c) = 3$ , it is sufficient to prove this:

**Statement.** *For any  $v_i \in V(K_n)$  and edges  $e = v_j v_i$  and  $f = v_i v_k$  such that  $c(e) < c(f)$ , there exists*

(Sa) *an edge  $g = v_j v_{j'}$ ,  $j' \notin \{i, j, k\}$ , such that  $c(g) < c(e)$ , or*

(Sb) *an edge  $g = v_k v_{k'}$ ,  $k' \notin \{i, j, k\}$ , such that  $c(f) < c(g)$ .*

Hence suppose there exist indices  $i, j, k \in I = \{0, \dots, n - 1\}$  such that for edges  $e = v_j v_i$  and  $f = v_i v_k$ , we have  $c(e) < c(f)$ , but neither (Sa) nor (Sb) holds. Then

$$c(v_j v_{j'}) > c(e) \quad \text{for all } j' \in I - \{i, j, k\}, \tag{1}$$

and

$$c(v_k v_{k'}) < c(f) \quad \text{for all } k' \in I - \{i, j, k\}. \tag{2}$$

We consider three cases, depending on the values of  $i$  and  $j$ .

Case 1:  $j \leq p - 1$ . Then, regardless of the values of  $i$  and  $j'$ ,  $c(v_j v_{j'}) = j + j'$  and  $c(e) = i + j$ . By (1),  $j' > i$  for all  $j' \in I - \{i, j, k\}$ . Hence  $i \leq 2$ . But  $p \geq 3$  since  $n \geq 6$ , and therefore  $i \leq p - 1$ . Now  $i + j = c(e) < c(f) = i + k$  implies that  $j < k$ . Therefore one of the following three subcases holds:

- (i)  $j = 0, k = 1$  and  $i = 2$ ,
- (ii)  $j = 0$  and  $k > i = 1$ ,
- (iii)  $i = 0$  and  $k > j > 0$ .

If (i) holds, then  $c(v_j v_k) = 1$ . Since  $n \geq 6$ , there exists  $k' \in I - \{0, 1, 2\}$  such that  $c(v_k v_{k'}) = k + k' \geq k' + 1 \geq 4 > c(f) = i + k = 3$ , contradicting (2). If (ii) holds, then  $c(f) = 1 + k$ . If  $k \leq p - 1$ , then  $v_k$  is adjacent to  $v_p$ , where  $p \notin \{0, 1, k\}$ , and  $c(v_k v_p) = k + p > c(f)$ , contradicting (2); while if  $k \geq p$ , then  $v_k$  is adjacent to  $v_2$  and  $c(v_2 v_k) = k + 2 > c(f)$ , again a contradiction. If (iii) holds, then  $c(e) = j < k = c(f)$ . If  $k \leq p - 1$ , then  $j < p - 1$  and  $v_k$  is adjacent to  $v_p$ , where  $p \notin \{0, j, k\}$ , giving a contradiction as in (ii). If  $k \geq p$ , then there exists  $\ell \in \{1, 2\} - \{j\}$  such that  $c(v_k v_\ell) = k + \ell > k$ , once again a contradiction.

Case 2:  $j \geq p$  and  $i \leq p - 1$ . Then  $c(e) = i + j$ . Since  $i \leq p - 1$  and  $n \geq 6$ , there exists  $j' \in I - \{i, j, k\}$  such that  $j' \geq p$ . Then  $c(v_{j'}v_j) = j + j' - 2p > i + j$  by (1); that is,  $i < j' - 2p \leq 0$ , which is impossible.

Case 3:  $\min\{i, j\} \geq p$ . Then  $c(e) = i + j - 2p$ . Suppose there exists  $j' \in I - \{i, j, k\}$  such that  $j' \geq p$ . Then  $c(v_{j'}v_j) = j + j' - 2p$  and thus  $j' > i$  by (1). Since  $i, j' \geq p$ ,

$$c(f) = c(v_iv_k) = \begin{cases} i + k & \text{if } k \leq p - 1, \\ i + k - 2p & \text{if } k \geq p, \end{cases}$$

and

$$c(v_kv_{j'}) = \begin{cases} j' + k & \text{if } k \leq p - 1, \\ j' + k - 2p & \text{if } k \geq p. \end{cases}$$

Thus, regardless of the value of  $k$ ,  $c(v_kv_{j'}) > c(f)$ . Since  $j' \in I - \{i, j, k\}$ , this contradicts (2). Hence there does not exist  $j' \in I - \{i, j, k\}$  such that  $j' \geq p$ . Since  $n \geq 6$ , we have  $|\{p, \dots, n - 1\}| \geq 3$ . We deduce that  $n \in \{6, 7\}$  and  $\{p, \dots, n - 1\} = \{i, j, k\}$  so that  $c(e) = i + j - 2p$  and  $c(f) = i + k - 2p$ , where  $j < k$  since  $c(e) < c(f)$ . For either value of  $n$ ,  $c(f) \leq 3$  and  $k \geq 4$ . Let  $j' = 0 < p$ . Then  $j' \in I - \{i, j, k\}$  and  $c(v_{j'}v_k) = j' + k = k \geq 4 > 3 \geq c(f)$ , again contradicting (2).  $\square$

The following corollary to Lemma 2 and Theorem 3 improves Theorem 17 of [Schurch 2013b].

**Corollary 4.** *For  $n \geq 6$ , we have  $\chi_\varepsilon(K_n) \leq \lfloor (3n - 3)/2 \rfloor$ .*

Combining Theorem 1 and Corollary 4 we improve Proposition 20 of [Schurch 2013b] and also obtain the new value  $\chi_\varepsilon(K_7)$ .

**Corollary 5.**  $\chi_\varepsilon(K_6) = 7$  and  $\chi_\varepsilon(K_7) = 9$ .

**3.2. The case  $n \equiv 0 \pmod{4}$ .** Our next result is an improved upper bound for  $\chi_\varepsilon(K_n)$  in the case where  $n \equiv 0 \pmod{4}$  and  $n \geq 8$ . Say  $n = 4m$  and  $V(K_n) = \{u_0, \dots, u_{2m-1}, v_0, \dots, v_{2m-1}\}$ . Let  $G$  and  $H$  be the subgraphs of  $K_n$  induced by  $\{u_0, \dots, u_{2m-1}\}$  and  $\{v_0, \dots, v_{2m-1}\}$ , respectively. Then  $G \cong H \cong K_{2m}$  and each of them is  $(2m - 1)$ -edge colourable. We describe a colouring  $c_1$  of  $K_n$  in the colours  $1, \dots, 4m + 1$  as follows.

- In  $G$ , let  $c_1$  be any proper edge colouring of  $K_{2m}$  in the  $2m - 1$  colours  $\{1, 2\} \cup \{m + 3, \dots, 3m - 1\}$ .
- In  $H$ , let  $c_1$  be any proper edge colouring of  $K_{2m}$  in the  $2m - 1$  colours  $\{4m, 4m + 1\} \cup \{m + 3, \dots, 3m - 1\}$ .
- We still need to colour the edges of the complete bipartite graph  $F \cong K_{2m, 2m}$  induced by the edges  $u_iv_j$ , with  $i, j \in \{0, \dots, 2m - 1\}$ . But  $\chi'(K_{2m, 2m}) = 2m$  and there are  $2m$  unused colours  $3, \dots, m + 2$  and  $3m, \dots, 4m - 1$ . Colour the edges of  $F$  with these colours.

It is clear that  $c_1$  is a proper edge colouring of  $K_{4m}$  in  $4m + 1$  colours.

**Theorem 6.** *For all  $m \geq 2$ , the colouring  $c_1$  of  $K_{4m}$  has flatness equal to three.*

*Proof.* Let  $F, G$  and  $H$  be the subgraphs of  $K_{4m}$  defined above and let  $e, f \in E(K_{4m})$  be adjacent edges such that  $c_1(e) < c_1(f)$ . We show that (Sa) or (Sb) holds, as stated in the proof of Theorem 3. We consider three cases, depending on the choice of  $e$  and  $f$ .

Case 1:  $\{e, f\} \cap E(F) = \emptyset$ . Assume first  $e, f \in E(G)$ ; say  $e = u_j u_i$  and  $f = u_i u_k$ . Then  $c_1(e) < c_1(f) \leq 3m - 1$ , and  $u_k$  is adjacent to some vertex  $v_\ell \in V(H)$  such that  $c_1(u_k v_\ell) = 4m - 1 > c_1(f)$ . Hence (Sb) holds. Similarly, if  $e, f \in E(H)$ , say  $e = v_j v_i$  and  $f = v_i v_k$ , then  $c_1(f) > c_1(e) \geq m + 3$ , and  $v_j$  is adjacent to some vertex  $u_\ell \in V(G)$  such that  $c_1(v_j u_\ell) = 3 < c_1(e)$ . Hence (Sa) holds.

Case 2:  $|\{e, f\} \cap E(F)| = 1$ . By symmetry we may assume that  $e \in E(F)$ ; say  $e = u_i v_j$ . If  $f \in E(G)$ , say  $f = u_i u_k$ , then  $c_1(e) \in \{3, \dots, m + 2\}$  and  $c_1(f) \in \{m + 3, m + 4, \dots, 3m - 1\}$ . Since  $m \geq 2$ ,  $u_k$  is adjacent to at least two vertices  $v_{t_1}, v_{t_2}$  of  $H$  such that  $c_1(u_k v_{t_\ell}) \in \{3m, \dots, 4m - 1\}$  for  $\ell = 1, 2$ , and we may choose a subscript  $t_\ell$ , say  $t_1$ , such that  $t_1 \neq j$ . Then  $v_j, u_i, u_k, v_{t_1}$  is a  $c_1$ -ascent of length three and (Sb) holds. On the other hand, if  $f \in E(H)$ , say  $f = v_j v_k$ , then  $c_1(e) \geq 3$ . In this case  $u_i$  is adjacent to a vertex  $u_\ell$  such that  $c_1(u_\ell u_i) \in \{1, 2\}$  and (Sa) holds.

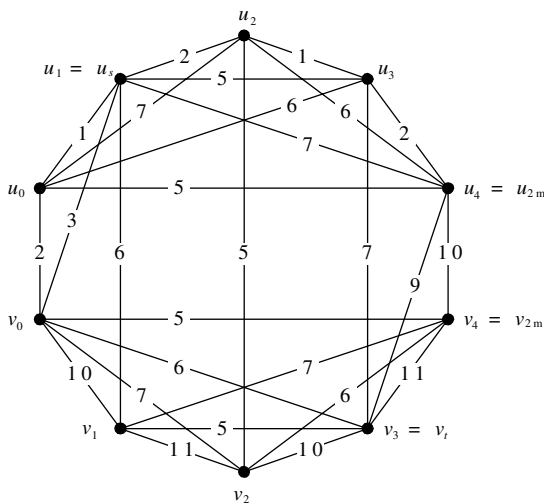
Case 3:  $\{e, f\} \subseteq E(F)$ . First, if  $e = u_i v_j$  and  $f = v_j u_k$ , then there exists at least one index  $\ell \in \{0, \dots, 2m - 1\} - \{i, k\}$  such that  $c_1(u_\ell u_i) \in \{1, 2\}$ . Then  $u_\ell, u_i, v_j, u_k$  is a  $c_1$ -ascent of length three and (Sa) holds. Finally, if  $e = v_i u_j$  and  $f = u_j v_k$ , then there exists at least one index  $\ell \in \{0, \dots, 2m - 1\} - \{i, k\}$  such that  $c_1(v_k v_\ell) \in \{4m, 4m + 1\}$ . Then  $v_i, u_j, v_k, v_\ell$  is a  $c_1$ -ascent of length three and (Sb) holds.  $\square$

Combining Theorems 1 and 6 we narrow down  $\chi_\varepsilon(K_n)$  to two possible values in infinitely many cases.

**Corollary 7.** *For all  $n \geq 8$  and  $n \equiv 0 \pmod{4}$ , we have  $n \leq \chi_\varepsilon(K_n) \leq n + 1$ .*

**3.3. The case  $n \equiv 2 \pmod{4}$ .** We now assume that  $n \equiv 2 \pmod{4}$  and  $n \geq 10$ . Say  $n = 4m + 2$  and  $V(K_n) = \{u_0, \dots, u_{2m}, v_0, \dots, v_{2m}\}$ . Let  $G$  and  $H$  be the subgraphs of  $K_n$  induced by  $\{u_0, \dots, u_{2m}\}$  and  $\{v_0, \dots, v_{2m}\}$ , respectively. Then  $G \cong H \cong K_{2m+1}$  and each of them is  $(2m+1)$ -edge colourable. We describe an edge colouring  $c_2$  of  $K_n$  in the colours  $1, \dots, 4m + 3$ . This colouring is similar to the colouring  $c_1$  above, but not quite as straightforward. See Figure 2 for a partial colouring of  $K_{10}$ .

- In  $G$ , let  $c_2$  be any proper edge colouring of  $K_{2m+1}$  in the  $2m + 1$  colours  $\{1, 2\} \cup \{m + 3, \dots, 3m + 1\}$ .
- In  $H$ , let  $c_2$  be any proper edge colouring of  $K_{2m+1}$  in the  $2m + 1$  colours  $\{4m + 2, 4m + 3\} \cup \{m + 3, \dots, 3m + 1\}$ .



**Figure 2.** Part of the edge colouring  $c_2$  of  $K_{10}$ .

We still need to colour the edges of the complete bipartite graph  $F \cong K_{2m+1,2m+1}$  induced by the edges  $u_i v_j$ , with  $i, j \in \{0, \dots, 2m\}$ . By König’s theorem,  $F$  is 1-factorable. Note that for each colour  $k$  in the edge colouring of  $G$  there is exactly one vertex that is not incident with an edge coloured  $k$ , and conversely, for each vertex  $u_i$  there is exactly one colour that does not occur as colour of an edge incident with  $u_i$ . A similar remark holds for  $H$ . Without loss of generality, say colour 2 does not occur at  $u_0$ , colour 1 does not appear at  $u_{2m}$ , colour  $4m + 3$  does not appear at  $v_0$  and colour  $4m + 2$  does not appear at  $v_{2m}$ . Since colour 2 does not occur at  $u_0$ , all other colours of the colouring do and thus there exists a vertex  $u_s \in V(G)$  such that  $c_2(u_0 u_s) = 1$ . Since colour  $4m + 2$  does not appear at  $v_{2m}$ , there exists a vertex  $v_t \in V(H)$  such that  $c_2(v_{2m} v_t) = 4m + 3$ .

- Colour the edges  $u_0 v_0$  and  $u_{2m} v_{2m}$  of  $F$  with colours 2 and  $4m + 2$ , respectively. For  $i, j \in \{1, \dots, 2m - 1\}$  and  $k \in \{m + 3, \dots, 3m + 1\}$ , colour  $u_i v_j$  with colour  $k$  if and only if no edge incident with  $u_i$  in  $G$  or with  $v_j$  in  $H$  is coloured  $k$ .

We have now coloured a 1-factor  $F_0$  of  $F$ , and  $F - F_0$  is a  $2m$ -regular bipartite graph, which is 1-factorable by König’s theorem. Let  $F'_1$  be a 1-factor of  $F - F_0$  that contains the edge  $v_0 u_s$ . If  $u_{2m} v_t \notin F'_1$ , let  $F_1 = F'_1$ , and if  $u_{2m} v_t \in F'_1$ , let  $u_i v_j \in F'_1 - \{v_0 u_s, u_{2m} v_t\}$  and define  $F_1 = (F'_1 - \{u_i v_j, u_{2m} v_t\}) \cup \{u_i v_t, u_{2m} v_j\}$ . Now  $F - F_0 - F_1$  is 1-factorable. Let  $F_2$  be a 1-factor of  $F - F_0 - F_1$  that contains  $u_{2m} v_t$ .

- Colour the edges in  $F_1$  with colour 3 and the edges in  $F_2$  with colour  $4m + 1$ . Colouring  $F - F_0 - F_1 - F_2$  with the  $2m - 2$  unused colours  $4, \dots, m + 2$  and  $3m + 2, \dots, 4m$  yields a proper edge colouring of  $K_{4m+2}$ .

**Theorem 8.** *For all  $m \geq 2$ , the colouring  $c_2$  of  $K_{4m+2}$  has flatness equal to three.*

*Proof.* Let  $F, J, G$  and  $H$  be the subgraphs of  $K_{4m+2}$  defined above and let  $e, f \in E(K_{4m+2})$  be adjacent edges such that  $c_2(e) < c_2(f)$ . We show that (Sa) or (Sb) holds, as stated in the proof of Theorem 3. If  $\{e, f\} \cap E(F) = \emptyset$ , the proof follows similar to Case 1 in the proof of Theorem 6. We consider two further cases.

Case 1:  $|\{e, f\} \cap E(F)| = 1$ . By symmetry we may assume that  $e \in E(F)$ ; say  $e = u_i v_j$ . First suppose that  $f \in E(G)$ , say  $f = u_i u_k$ . Since  $c_2(f) > c_2(e) \geq 2$ ,  $c_2(f) \in \{m + 3, \dots, 3m + 1\}$ . As in Case 2 of the proof of Theorem 3, (Sb) holds. Now suppose  $f = v_j v_k \in E(H)$ . If  $c_2(e) = 2$ , then  $i = j = 0$  and  $c_2(u_0 u_s) = 1$ . If  $c_2(e) \neq 2$  then  $c_2(e) > 2$  and there exists an index  $\ell$  such that  $c_2(u_i u_\ell) \in \{1, 2\}$ . Thus  $u_s, u_i, v_j, v_k$  or  $u_\ell, u_i, v_j, v_k$  is a  $c_2$ -ascent of length three and (Sa) holds.

Case 2:  $\{e, f\} \subseteq E(F)$ . Suppose  $e = u_i v_j$  and  $f = v_j u_k$ . If  $e = u_0 v_0$  and  $f = v_0 u_s$ , then  $c_2(e) = 2$  and  $c_2(f) = 3$ . Therefore there exists a vertex  $u_\ell$  such that  $c_2(u_s u_\ell) \in \{m + 3, \dots, 3m + 1\}$  and (Sb) holds. If  $e = u_0 v_0$  and  $k \neq s$ , then  $u_s, u_0, v_0, v_k$  is a  $c_2$ -ascent of length three and (Sa) holds. For all other choices of  $e = u_i v_j$  and  $f = v_j u_k$  it follows as in Case 3 of the proof of Theorem 3 that (Sa) or (Sb) holds. Suppose  $e = v_i u_j$  and  $f = u_j v_k$ . If  $e = v_t u_{2m}$  and  $f = u_{2m} v_{2m}$ , then  $c_2(e) = 4m + 1$  and  $c_2(f) = 4m + 2$ . There exists a vertex  $v_\ell$  such that  $c_2(v_\ell v_t) \in \{m + 3, \dots, 3m + 1\}$  and thus (Sa) holds. If  $f = u_{2m} v_{2m}$  and  $i \neq t$ , then  $v_i, v_{2m}, u_{2m}, v_t$  is a  $c_2$ -ascent of length three and (Sb) holds. All other cases are dealt with as in Case 3 of the proof of Theorem 3.  $\square$

Combining Theorems 1 and 8 and Corollary 4 determines  $\chi_\varepsilon(K_n)$  for all  $n \equiv 2 \pmod{10}$ ,  $n \geq 6$ .

**Corollary 9.** *For all  $n \geq 6$  and  $n \equiv 2 \pmod{10}$ , we have  $\chi_\varepsilon(K_n) = n + 1$ .*

#### 4. Conclusion

In Theorem 1 we proved a lower bound for  $\chi_\varepsilon(K_n)$ , and in Corollary 4 we improved the previously known general upper bound for  $\chi_\varepsilon(K_n)$  from  $2n - 3$  to  $\lfloor (3n - 3)/2 \rfloor$ . Corollary 7 improves this bound for  $n \equiv 0 \pmod{4}$  and allows us to bound  $\chi_\varepsilon(K_{4m})$  by  $4m \leq \chi_\varepsilon(K_n) \leq 4m + 1$ . Finally, Corollary 9 determines  $\chi_\varepsilon(K_n)$  for all  $n \equiv 2 \pmod{4}$ ,  $n \geq 6$ . Based on the results for even  $n$  and the values  $\chi_\varepsilon(K_5) = 7$  and  $\chi_\varepsilon(K_7) = 9$ , we formulate the following conjecture.

**Conjecture 10.** *For all  $n \geq 4$ , we have  $\chi_\varepsilon(K_n) = \chi'(K_n) + 2$ .*

#### Acknowledgements

Jean Breytenbach wishes to thank Professor Jan van Vuuren of the Department of Industrial Engineering, Stellenbosch University, for fuelling his interest in graph the-

ory. Both authors hereby also express their gratitude towards Professor van Vuuren for introducing them and providing a wonderful research environment to work in.

## References

- [Bialostocki and Roditty 1987] A. Bialostocki and Y. Roditty, “A monotone path in an edge-ordered graph”, *Internat. J. Math. Math. Sci.* **10**:2 (1987), 315–320. MR 88b:05087 Zbl 0633.05043
- [Burger et al. 2005] A. P. Burger, E. J. Cockayne, and C. M. Mynhardt, “Altitude of small complete and complete bipartite graphs”, *Australas. J. Combin.* **31** (2005), 167–177. MR 2005i:05160 Zbl 1080.05046
- [Calderbank et al. 1984] A. R. Calderbank, F. R. K. Chung, and D. G. Sturtevant, “Increasing sequences with nonzero block sums and increasing paths in edge-ordered graphs”, *Discrete Math.* **50**:1 (1984), 15–28. MR 85k:05062 Zbl 0542.05058
- [Chartrand et al. 2011] G. Chartrand, L. Lesniak, and P. Zhang, *Graphs & digraphs*, 5th ed., CRC Press, Boca Raton, FL, 2011. MR 2012c:05001 Zbl 1211.05001
- [Chvátal and Komlós 1971] V. Chvátal and J. Komlós, “Some combinatorial theorems on monotonicity”, *Canad. Math. Bull.* **14** (1971), 151–157. MR 49 #2445 Zbl 0214.23503
- [Cockayne and Mynhardt 2006] E. J. Cockayne and C. M. Mynhardt, “A lower bound for the depression of trees”, *Australas. J. Combin.* **35** (2006), 319–328. MR 2007j:05118 Zbl 1094.05018
- [Cockayne et al. 2006] E. J. Cockayne, G. Geldenhuys, P. J. P. Grobler, C. M. Mynhardt, and J. H. van Vuuren, “The depression of a graph”, *Util. Math.* **69** (2006), 143–160. Preprint (2004) available at <http://tinyurl.com/GraphDepression2004>.
- [Gaber-Rosenblum and Roditty 2009] I. Gaber-Rosenblum and Y. Roditty, “The depression of a graph and the diameter of its line graph”, *Discrete Math.* **309**:6 (2009), 1774–1778. MR 2010d:05045 Zbl 1205.05066
- [Graham and Kleitman 1973] R. L. Graham and D. J. Kleitman, “Increasing paths in edge ordered graphs”, *Period. Math. Hungar.* **3** (1973), 141–148. MR 48 #5910 Zbl 0243.05116
- [Mynhardt 2008] C. M. Mynhardt, “Trees with depression three”, *Discrete Math.* **308**:5-6 (2008), 855–864. MR 2008j:05186 Zbl 1149.05042
- [Mynhardt and Schurch 2013] C. M. Mynhardt and M. Schurch, “A class of graphs with depression three”, *Discrete Math.* **313**:11 (2013), 1224–1232. MR 3034754 Zbl 1277.05083
- [Mynhardt and Schurch 2014] C. M. Mynhardt and M. Schurch, “A construction of a class of graphs with depression three”, *Australas. J. Combin.* **58**:2 (2014), 249–263. Zbl 1296.05087
- [Mynhardt et al. 2005] C. M. Mynhardt, A. P. Burger, T. C. Clark, B. Falvai, and N. D. R. Henderson, “Altitude of regular graphs with girth at least five”, *Discrete Math.* **294**:3 (2005), 241–257. MR 2006a:05079 Zbl 1062.05131
- [Roditty et al. 2001] Y. Roditty, B. Shoham, and R. Yuster, “Monotone paths in edge-ordered sparse graphs”, *Discrete Math.* **226**:1-3 (2001), 411–417. MR 2001i:05096 Zbl 0961.05040
- [Schurch 2013a] M. Schurch, *On the depression of graphs*, Ph.D. thesis, University of Victoria, 2013, available at <https://dspace.library.uvic.ca:8443/handle/1828/4527>.
- [Schurch 2013b] M. Schurch, “Edge colourings and the depression of a graph”, *J. Combin. Math. Combin. Comput.* **85** (2013), 195–212. MR 3088160 Zbl 1274.05176
- [Schurch and Mynhardt 2014] M. Schurch and C. Mynhardt, “The depression of a graph and  $k$ -kernels”, *Discuss. Math. Graph Theory* **34**:2 (2014), 233–247. MR 3194034 Zbl 1290.05128

[Yuster 2001] R. Yuster, “Large monotone paths in graphs with bounded degree”, *Graphs Combin.* **17**:3 (2001), 579–587. MR 2002k:05135 Zbl 1010.05044

Received: 2013-07-22 Accepted: 2013-10-26

`jabreytenbach@cs.sun.ac.za` *Computer Science, Department of Mathematical Sciences,  
Stellenbosch University, Private Bag X1, Matieland,  
7602, South Africa*

`kieka@uvic.ca` *Department of Mathematics and Statistics, University of Victo-  
ria, P.O. Box 1700 STN CSC, Victoria, BC V8W 2Y2, Canada*





# Bisection envelopes

Noah Fechter-Pradines

(Communicated by Frank Morgan)

We study the envelope of the family of lines which bisect the interior region of a simple, closed curve in the plane. We determine this bisection envelope for polygons and show that polygons with no parallel pairs of sides are characterized by their bisection envelope. We show that the bisection envelope always has at least three and an odd number of cusps. We investigate the winding numbers of bisection envelopes, and use this to show that there are an infinite number of curves with any given bisection envelope and show how to generate them. We obtain results on the intersections of bisecting lines. Finally, we give a relationship between the internal area of a curve and that of its bisection envelope.

## 1. Introduction and overview

We study the envelope of the family of lines that bisect the interior region of a given simple, closed curve in the plane. This concept, which we call the bisection envelope, was explored in [Fusco and Pratelli 2011]; however, here we apply it to a more general class of curves. Fusco and Pratelli only used the bisection envelope in relation to Zindler sets — convex sets whose bisecting chords have fixed length: they used as a tool to rewrite the problem of minimizing the area of a Zindler set with fixed bisecting chord length.

Specifically, let  $\mathcal{S}$  be a simple compact curve which is piecewise of class  $C^1$  with a finite number of pieces. Let  $\mathcal{L}$  be the set of lines  $l_\theta$  that have direction  $\theta$  and bisect the interior of  $\mathcal{S}$ . The bisection envelope of  $\mathcal{S}$  is the envelope of the lines in  $\mathcal{L}$ . For curves  $\mathcal{S}$  that are bisection convex (see Definition 2.2), we show that the bisection envelope is the midpoint locus of bisecting chords. Furthermore, we show that for curves that are strictly bisection convex (see Definition 2.3) we can parametrize the bisection envelope by a function  $f$  such that  $f(\theta)$  lies on  $l_\theta$ , and find the derivative of  $f$ , defined at all but a finite number of points. Where

---

*MSC2010:* 26B15, 51M25.

*Keywords:* bisection envelope, area, winding number, envelope, geometry, bisection.

At the time of the writing, the author was a student at the British International School of Boston. He is now an undergraduate at Harvard University.

this derivative exists we show it is of the form  $v_\theta(\cos \theta, \sin \theta)$ , and give conditions on  $\mathcal{S}$  such that for a scalar  $v_\theta$ ,

$$f(\theta) = f(0) + \int_0^\theta v_t(\cos t, \sin t) dt.$$

We show that zeros of  $f'(\theta)$  (which are also zeros of  $v_\theta$ ) each corresponds to a bisecting chord at whose endpoints the tangents to  $\mathcal{S}$  are parallel. We also show a relation between sign changes of  $v_\theta$  and the appearance of cusps on the bisection envelope. These results are summarized in Theorem 1.

In Section 3, we examine the bisection envelopes of polygons, showing that they are the union of sections of hyperbolas. Furthermore, for each hyperbola, there exist two sides of  $\mathcal{S}$  which are segments of its asymptotes. We also show Theorem 2, which states that polygons with no mutually tangent sides are uniquely defined by their bisection envelopes.

Section 4 addresses curves with identical bisection envelopes. We show how to generate a curve  $\mathcal{S}'$  from the bisection envelope  $\mathcal{B}$  of a strictly bisection convex curve  $\mathcal{S}$  satisfying certain criteria by letting  $\mathcal{S}'$  be the image of a function  $g$ , defined as

$$g(\theta) = f(\theta) + r(\theta)(\cos \theta, \sin \theta),$$

where  $r(\theta)$  is a radius function that can be changed to produce different  $\mathcal{S}'$ . The main result of Section 4 is Theorem 3, which states that if the generated  $\mathcal{S}'$  does not intersect  $\mathcal{B}$ , then  $\mathcal{B}$  is indeed the bisection envelope of both  $\mathcal{S}$  and  $\mathcal{S}'$ .

To prove Theorem 3, we first prove Theorem 4, which concerns the winding numbers of bisection envelopes. Specifically, let  $m_P$  be the number of lines through a point  $P$  tangent to  $\mathcal{B}$ . We show that

$$m_P = -2w(P) + 1,$$

where  $w(P)$  is the winding number of  $\mathcal{B}$  about  $P$  with  $\theta$  increasing from 0 to  $\pi$ .

In Section 5, we examine the interior areas of  $\mathcal{S}'$  and  $\mathcal{B}$ . The interior area of  $\mathcal{B}$  is usually not well-defined, as it can be self-intersecting, therefore we define the interior area of a curve  $\Gamma$  by the integral

$$\mathcal{A}(\Gamma) = \frac{1}{2} \oint_{\Gamma} x dy - y dx.$$

From this definition, we use the construction in Section 4 to break apart  $\mathcal{A}(\mathcal{S}')$  to give Theorem 5, which states that

$$\mathcal{A}(\mathcal{S}') = \int_0^{2\pi} \frac{r^2(\theta)}{2} d\theta + 2\mathcal{A}(\mathcal{B}).$$

We also show that  $\mathcal{A}(\mathcal{B})$  is never positive and use this to show that certain curves with maximal interior area are rotationally symmetric (see Corollary 5.3). We

conclude by computing the internal area of the bisection envelope of an equilateral triangle, and thus deduce a constant universal to all triangles:  $\frac{3}{4} \ln 2 - \frac{1}{2}$ , the ratio of the area of a triangle to the area of its bisection envelope.

### 2. Basic properties

For the entirety of this paper, it is assumed that  $S$  is a curve in  $\mathbb{R}^2$  which is compact, continuous, simple, and piecewise of class  $C^1$  with a finite number of pieces.

We now define the bisection envelope.

**Definition 2.1.** Given such a curve  $S$ , define  $\mathcal{L}$  to be the family of lines that bisect the interior area of  $S$ . Each  $l_\theta \in \mathcal{L}$  is the bisecting line in direction  $\theta$ . Define the *bisection envelope*  $\mathcal{B}$  of  $S$  to be the envelope of  $\mathcal{L}$ ; that is,

$$\mathcal{B} = \left\{ P \mid P = \lim_{\epsilon \rightarrow 0} l_\theta \cap l_{\theta+\epsilon}, 0 \leq \theta < \pi \right\}.$$

We now restrict the class of curves  $S$  to be studied.

**Definition 2.2.** Define  $S$  and  $\mathcal{L}$  as above. We say that  $S$  is *bisection convex* if for all  $\theta$ ,  $l_\theta$  intersects  $S$  in exactly two points. Alternatively, for every point  $A$  on  $S$ , there exists a unique point  $B$  also on  $S$  such that the line  $AB$  bisects the interior area of  $S$ .

We also create a tighter restriction.

**Definition 2.3.** Define  $S$  and  $\mathcal{L}$  as before. We say that  $S$  is *strictly bisection convex* if it is bisection convex and for all  $\theta$ ,  $l_\theta$  is not tangent to  $S$ . At any point where there are two tangents to  $S$  — one from each side — the  $l_\theta$  through that point is distinct from both tangents.

Henceforth, unless otherwise stated, *it is assumed that  $S$  is strictly bisection convex.*

Define  $A(\theta)$  and  $B(\theta)$  to be the endpoints of the bisecting chord in direction  $\theta$ , with  $B(\theta) = A(\theta + \pi)$ . We distinguish between  $A(\theta)$  and  $B(\theta)$  by demanding that for each point  $Q \neq A(\theta)$ ,  $B(\theta)$  on the bisecting chord, the vector  $A(\theta) - Q$  points in positive direction  $\theta$  and the vector  $B(\theta) - Q$  points in positive direction  $\theta + \pi$ .

**Proposition 2.4.** *Assume that  $S$  is bisection convex. Then  $A(\theta)$  varies continuously with  $\theta$ .*

*Proof.* First, we note that any two bisecting chords must intersect in the interior of  $S$ , for if they did not, the interior of  $S$  would be split into three regions, one of which would have zero area, which does not make sense.

From this, we have  $\lim_{\epsilon \rightarrow 0} l_{\theta+\epsilon} = l_\theta$ , as the limit of the intersection point  $l_{\theta+\epsilon} \cap l_\theta$  is bounded. This also implies that the limit as  $\epsilon \rightarrow 0$  of the distance from  $A(\theta + \epsilon)$  to the intersection point  $l_{\theta+\epsilon} \cap l_\theta$  is bounded. Therefore, the limit as  $\epsilon \rightarrow 0$  of the perpendicular distance from  $A(\theta + \epsilon)$  to  $l_\theta$  is zero.

We have that  $\lim_{\epsilon \rightarrow 0} A(\theta + \epsilon)$  must be a point  $P$  on  $l_\theta$  which intersects  $\mathcal{S}$ , where for every other point  $Q$  on the bisecting chord with direction  $\theta$ , the vector  $P - Q$  points in positive direction  $\theta$ . There is only one such point,  $A(\theta)$ ; therefore,

$$\lim_{\epsilon \rightarrow 0} A(\theta + \epsilon) = A(\theta),$$

and  $A(\theta)$  varies continuously with  $\theta$ .  $\square$

From this,  $B(\theta)$  also varies continuously with  $\theta$ . We now determine the bisection envelope of bisection convex curves.

**Proposition 2.5.** *Let  $\mathcal{S}$  be bisection convex. Fix  $\theta$  and let  $A = A(\theta)$  and  $B = B(\theta)$ . Then,*

$$\lim_{\epsilon \rightarrow 0} l_\theta \cap l_{\theta+\epsilon} = \frac{A + B}{2}. \quad (2-1)$$

*Proof.* Let  $A(\theta + \epsilon) = A_\epsilon$  and  $B(\theta + \epsilon) = B_\epsilon$ . Let  $l_\theta \cap l_{\theta+\epsilon} = O_\epsilon$ , and let  $\lim_{\epsilon \rightarrow 0} l_\theta \cap l_{\theta+\epsilon} = O$ ; see Figure 1. Define  $a(\epsilon) = d(A_\epsilon, O_\epsilon)$ ,  $b(\epsilon) = d(B_\epsilon, O_\epsilon)$ , and extend to let  $a(0) = d(A, O)$  and  $b(0) = d(O, B)$ .

Since  $l_\theta, l_{\theta+\epsilon}$  are bisecting line segments,

$$\mathcal{A}(AO_\epsilon A_\epsilon) = \mathcal{A}(BO_\epsilon B_\epsilon), \quad (2-2)$$

where  $AO_\epsilon A_\epsilon$  and  $BO_\epsilon B_\epsilon$  are not triangles, but rather the regions enclosed by  $\mathcal{S}, l_\theta$ , and  $l_{\theta+\epsilon}$ .

For fixed  $\epsilon$ , we have the inequality

$$\frac{1}{2}\epsilon m^2 \leq \mathcal{A}(AO_\epsilon A_\epsilon) \leq \frac{1}{2}\epsilon M^2,$$

where  $m$  and  $M$  are the minimum and maximum values of  $d(A_\delta, O_\epsilon)$  for  $0 \leq \delta \leq \epsilon$ .

As  $m \leq a(\epsilon) \leq M$ ,

$$\frac{1}{2}\epsilon m^2 \leq \frac{1}{2}\epsilon a^2(\epsilon) \leq \frac{1}{2}\epsilon M^2.$$

The previous two inequalities have the same bounds, therefore

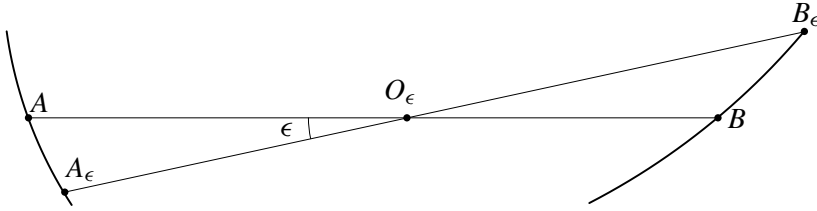
$$\left| \mathcal{A}(AO_\epsilon A_\epsilon) - \frac{1}{2}\epsilon a^2(\epsilon) \right| \leq \frac{1}{2}\epsilon(M^2 - m^2). \quad (2-3)$$

From the continuity of  $\mathcal{S}$ , we have

$$\lim_{\epsilon \rightarrow 0} \frac{\frac{1}{2}\epsilon(M^2 - m^2)}{\epsilon} = \frac{1}{2}(a^2(0) - a^2(0)) = 0.$$

Combining this with (2-3) and using an identical argument for  $\mathcal{A}(BO_\epsilon B_\epsilon)$ , we have

$$\begin{aligned} \left| \mathcal{A}(AO_\epsilon A_\epsilon) - \frac{1}{2}\epsilon a^2(\epsilon) \right| &= o(\epsilon), \\ \left| \mathcal{A}(BO_\epsilon B_\epsilon) - \frac{1}{2}\epsilon b^2(\epsilon) \right| &= o(\epsilon). \end{aligned} \quad (2-4)$$



**Figure 1.** The situation considered in the proof of Proposition 2.5.

By the triangle inequality and (2-2), we have that

$$\begin{aligned} \left| \frac{1}{2}\epsilon a^2(\epsilon) - \frac{1}{2}\epsilon b^2(\epsilon) \right| &\leq \left| \frac{1}{2}\epsilon a^2(\epsilon) - \mathcal{A}(A O_\epsilon A_\epsilon) \right| + \left| \mathcal{A}(A O_\epsilon A_\epsilon) - \frac{1}{2}\epsilon b^2(\epsilon) \right| \\ &= \left| \mathcal{A}(A O_\epsilon A_\epsilon) - \frac{1}{2}\epsilon a^2(\epsilon) \right| + \left| \mathcal{A}(B O_\epsilon B_\epsilon) - \frac{1}{2}\epsilon b^2(\epsilon) \right|. \end{aligned}$$

It follows from this and (2-4) that

$$\begin{aligned} \left| \frac{1}{2}\epsilon a^2(\epsilon) - \frac{1}{2}\epsilon b^2(\epsilon) \right| &= o(\epsilon), \\ \left| \frac{1}{2}a^2(0) - \frac{1}{2}b^2(0) \right| &= 0, \\ a(0) &= b(0). \end{aligned} \tag{2-5}$$

Therefore  $O$  is the midpoint of  $A$  and  $B$ . □

Hence,  $\mathcal{B}$  is the locus of midpoints of the intersections of each  $l_\theta \in \mathcal{L}$  with  $S$ .

Define a function  $f : \mathbb{R} \rightarrow \mathbb{R}^2$ , with  $f(\theta + \pi) = f(\theta)$ , such that  $f(\theta)$  signifies the point on  $\mathcal{B}$  that is the midpoint of the bisecting chord of  $S$  with direction  $\theta$ . The image of this function is  $\mathcal{B}$ . We are interested in the derivative of this function, where it exists.

**Proposition 2.6.** *Let  $S$  be strictly bisection convex. Fix  $\theta$  such that  $S$  is of class  $C^1$  at the endpoints  $A(\theta)$ ,  $B(\theta)$  of the bisecting chord with direction  $\theta$ . Then  $f'(\theta)$  is defined, and if  $f'(\theta)$  is nonzero, then  $l_\theta$  is tangent to  $\mathcal{B}$  at  $f(\theta)$ .*

*Proof.* It suffices to derive  $f'(\theta)$  and show that it is either zero or a nonzero vector pointing in direction  $\theta$ .

Without loss of generality, let the axes be redefined such that direction  $\theta$  is along the  $x$ -axis.

Define  $A$ ,  $B$ ,  $A_\epsilon$ ,  $B_\epsilon$ ,  $O_\epsilon$  as in the proof of Proposition 2.5. Let

$$M = \frac{A + B}{2} \quad \text{and} \quad M_\epsilon = \frac{A_\epsilon + B_\epsilon}{2}.$$

Let  $r = d(A, M) = d(M, B)$ ,  $r(\epsilon) = d(A_\epsilon, M_\epsilon) = d(M_\epsilon, B_\epsilon)$ ,  $\lambda(\epsilon) = d(O_\epsilon, M_\epsilon)$ . Let  $\alpha(\epsilon) = m\angle A_\epsilon A O_\epsilon$  and  $\beta(\epsilon) = m\angle B_\epsilon B O_\epsilon$ .

Let  $a_h(\epsilon)$  and  $a_v(\epsilon)$  be the horizontal and vertical components of  $\overrightarrow{AA_\epsilon}$ , positive in directions  $\theta$  and  $\theta + \pi/2$  respectively. Define  $b_h(\epsilon)$  and  $b_v(\epsilon)$  similarly; see Figure 2.

By definition,

$$\begin{aligned} f'(\theta) &= \lim_{\epsilon \rightarrow 0} \frac{\overrightarrow{MM_\epsilon}}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{\overrightarrow{AA_\epsilon} + \overrightarrow{BB_\epsilon}}{2\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} \left( \frac{a_h(\epsilon) + b_h(\epsilon)}{2\epsilon}, \frac{a_v(\epsilon) + b_v(\epsilon)}{2\epsilon} \right). \end{aligned} \quad (2-6)$$

By inspection,

$$a_v(\epsilon) = -(r(\epsilon) - \lambda(\epsilon)) \sin \epsilon \quad \text{and} \quad b_v(\epsilon) = (r + \lambda(\epsilon)) \sin \epsilon.$$

Thus

$$\lim_{\epsilon \rightarrow 0} \frac{a_v(\epsilon) + b_v(\epsilon)}{\epsilon} = \lim_{\epsilon \rightarrow 0} (r - r(\epsilon) + 2\lambda(\epsilon)) \frac{\sin \epsilon}{\epsilon} = 0, \quad (2-7)$$

as  $\lim_{\epsilon \rightarrow 0} r(\epsilon) = r$  and  $\lim_{\epsilon \rightarrow 0} M_\epsilon = \lim_{\epsilon \rightarrow 0} O_\epsilon = M$ , which follow from definition and Proposition 2.5.

As  $a_h(\epsilon) = -a_v(\epsilon) \cot(\alpha(\epsilon))$  and  $b_h(\epsilon) = -b_v(\epsilon) \cot(\beta(\epsilon))$ , we have

$$\begin{aligned} &\lim_{\epsilon \rightarrow 0} \frac{a_h(\epsilon) + b_h(\epsilon)}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} \left( r(\epsilon) \cot(\alpha(\epsilon)) - r \cot(\beta(\epsilon)) - \lambda(\epsilon) \cot(\beta(\epsilon)) - \lambda(\epsilon) \cot(\alpha(\epsilon)) \right) \frac{\sin \epsilon}{\epsilon} \\ &= \lim_{\epsilon \rightarrow 0} \left( r(\cot(\alpha(\epsilon)) - \cot(\beta(\epsilon))) - \cot(\alpha(\epsilon))(r - r(\epsilon)) - \lambda(\epsilon) \cot(\beta(\epsilon)) \right. \\ &\quad \left. - \lambda(\epsilon) \cot(\alpha(\epsilon)) \right) \frac{\sin \epsilon}{\epsilon} \\ &= r(\cot \alpha - \cot \beta), \quad \text{where } \alpha = \lim_{\epsilon \rightarrow 0} \alpha(\epsilon), \quad \beta = \lim_{\epsilon \rightarrow 0} \beta(\epsilon). \end{aligned} \quad (2-8)$$

This follows from the same limits stated earlier, as  $\mathcal{S}$  is strictly bisection convex and thus neither  $\alpha$  nor  $\beta$  are 0 or  $\pi$ . Note that  $\alpha$  and  $\beta$  are not necessarily defined—the limits only exist if  $\mathcal{S}$  is of class  $C^1$  locally at  $A$  and  $B$ , and thus  $\alpha$  and  $\beta$  are not defined for only a finite number of values of  $\theta$ . Where they are defined, we can combine (2-6), (2-7), and (2-8), giving

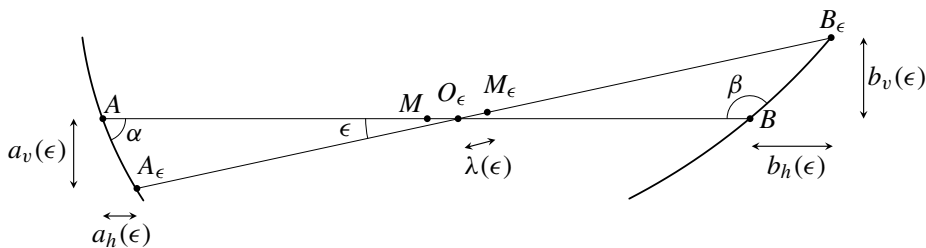
$$f'(\theta) = \left( \frac{r(\cot \alpha - \cot \beta)}{2}, 0 \right), \quad (2-9)$$

and so  $f'(\theta)$  is defined. Since  $f'(\theta)$  has  $y$ -component 0, it points in direction  $\theta$  if it is nonzero.  $\square$

Directly from (2-9), we have:

**Corollary 2.7.** *We have  $f'(\theta) = 0$  if and only if the tangents to  $\mathcal{S}$  at the endpoints of the bisecting chord with direction  $\theta$  are parallel, that is, when  $\alpha = \beta$ .*

Proposition 2.6 can be further extended to cover more points on  $\mathcal{B}$ .



**Figure 2.** The situation considered in the proof of Proposition 2.6.

**Proposition 2.8.** *If  $f'$  is zero or undefined at a finite number of points, then for all  $\theta$ ,  $l_\theta$  is tangent to  $\mathcal{B}$  at  $f(\theta)$ .*

*Proof.* Define  $t_\theta$  to be the tangent to  $\mathcal{B}$  at  $f(\theta)$ .

If there are only a finite number of points for which  $f'$  is zero or undefined, then there are only a finite number of values of  $\theta$  for which Proposition 2.6 does not hold. Thus, around any of these values  $\theta_0$ , there exists a neighborhood for which Proposition 2.6 does hold. For small  $\epsilon$ ,  $\theta_0 + \epsilon$  will lie in this neighborhood. Also,  $f$  is continuous, so the lines  $l_\theta \in \mathcal{L}$  vary continuously with  $\theta$ , and it is clear that

$$t_{\theta_0} = \lim_{\epsilon \rightarrow 0} t_{\theta_0 + \epsilon} = \lim_{\epsilon \rightarrow 0} l_{\theta_0 + \epsilon} = l_{\theta_0}. \quad \square$$

From the derivation in Proposition 2.6, it is true that wherever  $f'$  is defined, it points in direction  $\theta$ ; thus, each defined  $f'(\theta)$  is a scalar multiple of  $(\cos \theta, \sin \theta)$ .

Also from Proposition 2.6, we have:

**Proposition 2.9.** *Wherever  $f'(\theta)$  is defined,  $f'$  is continuous at  $\theta$ .*

*Proof.* From (2-9) we have that, where  $f'(\theta)$  is defined, it is continuous if  $r$ ,  $\cot \alpha$ , and  $\cot \beta$  vary continuously with  $\theta$ .

We have that  $r$  is half of the distance between the points  $A(\theta)$  and  $B(\theta)$ , which vary continuously by Proposition 2.4, and therefore varies continuously for any  $\theta$ .

From the fact that  $\mathcal{S}$  is strictly bisection convex, the angle  $\alpha$  must remain between 0 and  $\pi$ ; therefore,  $\cot \alpha$  varies continuously if  $\alpha$  varies continuously. The angle  $\alpha$  is defined as the difference in direction of the bisecting line and the direction of the tangent to  $\mathcal{S}$  at  $A(\theta)$ . The direction of the bisecting line is  $\theta$ , so it varies continuously. Where  $f(\theta)$  is defined,  $\mathcal{S}$  is of class  $C^1$  locally at  $A(\theta)$ , and as  $A(\theta)$  is a continuous parametrization of  $\mathcal{S}$ , the tangents to  $\mathcal{S}$  around  $A(\theta)$  vary continuously with  $\theta$ . Thus  $\alpha$  varies continuously with  $\theta$ .

An identical argument can be used to show that  $\beta$  varies continuously with  $\theta$ , and the result follows.  $\square$

From this, we have that  $f'$  is undefined in at most a finite number of places over any period of length  $2\pi$ , and it is only at these points that it is discontinuous.

**Definition 2.10.** Define  $v_\theta := f'(\theta) \cdot (\cos \theta, \sin \theta)$ , where  $f'$  is defined. Then  $v_\theta$  has the following properties:

- (1)  $|v_\theta| = |f'(\theta)|$ .
- (2)  $v_{\theta+\pi} = -v_\theta$ .
- (3)  $f'(\theta) = v_\theta(\cos \theta, \sin \theta)$ .
- (4)  $\int_{\theta_0}^{\theta_0+\pi} v_\theta(\cos \theta, \sin \theta) d\theta = (0, 0)$ .

These follow directly from the definition of  $v_\theta$  and from Proposition 2.6. Also note that the integral shown is defined, as the number of discontinuities of  $v_\theta$  over the interval is the same as the number of discontinuities of  $f'$ , thus finite, and the set of discontinuity points has measure 0.

**Proposition 2.11.** *If  $v_\theta$  is not identically zero, then over any interval  $[\theta_0, \theta_0 + \pi]$  where  $v_{\theta_0} \neq 0$ ,  $v_\theta$  changes sign an odd number of times, and at least thrice.*

*Proof.* As  $v_{\theta_0+\pi} = -v_{\theta_0}$ , we know that  $v_\theta$  must change sign at least once in the interval and must change an odd number of times.

Assume that only one sign change occurs over the interval  $[\theta_0, \theta_0 + \pi]$ . Then there exists a value  $\theta_1$  (not necessarily unique) with  $\theta_0 < \theta_1 < \theta_0 + \pi$  such that over the interval  $[\theta_0, \theta_1]$ ,  $v_\theta \leq 0$  and over the interval  $[\theta_1, \theta_0 + \pi]$ ,  $v_\theta \geq 0$ , or vice versa. Either way, this ensures that  $v_\theta$  does not change sign over the interval  $[\theta_1, \theta_1 + \pi]$ .

Consider the component of  $f'(\theta)$  in direction  $\theta_1 + \pi/2$ . We observe that

$$\begin{aligned} 0 &= \int_{\theta_1}^{\theta_1+\pi} f'(\theta) \cdot (\cos(\theta_1 + \pi/2), \sin(\theta_1 + \pi/2)) d\theta \\ &= \int_{\theta_1}^{\theta_1+\pi} v_\theta(\cos \theta, \sin \theta) \cdot (-\sin(\theta_1), \cos(\theta_1)) d\theta \\ &= \int_{\theta_1}^{\theta_1+\pi} v_\theta \sin(\theta - \theta_1) d\theta. \end{aligned} \tag{2-10}$$

Neither  $v_\theta$  nor  $\sin(\theta - \theta_1)$  change sign between the bounds of the integral; thus, their product does not change sign (and is not identically zero by assumption), and (2-10) cannot be equal to 0, a contradiction.

This implies there is more than one sign change in any such interval  $[\theta_0, \theta_0 + \pi]$ , so there are at least three, the next odd number.  $\square$

**Remark 2.12.** The notion of sign changes of  $v_\theta$  has a geometric manifestation. For every point or interval where  $v_\theta$  changes sign, a cusp or corner, respectively, appears on  $\mathcal{B}$ . If  $v_\theta$  is zero at a finite number of points, then corners do not occur, and we have one cusp per sign change in an interval of length  $\pi$ . With these conditions, we extend Proposition 2.11 to  $\mathcal{B}$  geometrically — if  $\mathcal{B}$  is not a point and has no corners, then it has an odd number of cusps, and at least three cusps.



Note that this collection of results becomes much cleaner if we assume  $\mathcal{S}$  to be entirely of class  $C^1$ .

**Theorem 1.** *If  $\mathcal{S}$  is strictly bisection convex and of class  $C^1$ , then there exist  $n \geq 3$  lines  $l_\theta$  that bisect the interior area of  $\mathcal{S}$  such that the tangents to  $\mathcal{S}$  at  $A(\theta)$  and  $B(\theta)$  are parallel. If  $n$  is finite, then there exist  $m$  cusps on the bisection envelope  $\mathcal{B}$  of  $\mathcal{S}$ , with  $n \geq m \geq 3$  and  $m$  odd.*

*Proof.* From our assumptions and Propositions 2.6 and 2.9,  $f'$  is defined everywhere and is continuous; therefore, from the definition of  $v_\theta$ , we know that  $v_\theta$  is continuous. Therefore, if we let  $m$  be the number of sign changes of  $v_\theta$  and  $n$  be the number of zeros, we have  $n \geq m$ . A zero of  $v_\theta$  is a zero of  $f(\theta)$ , and thus by Corollary 2.7, there are  $n$  lines  $l_\theta$  such that the tangents to  $\mathcal{S}$  at  $A(\theta)$  and  $B(\theta)$  are parallel. If  $n$  is finite, then  $v_\theta$  is not identically zero, so by Proposition 2.11,  $m$  is odd and at least 3. With  $n$  finite, no corners exist on  $\mathcal{B}$ , so from Remark 2.12, we have that  $m$  is the number of cusps on  $\mathcal{B}$ . □

### 3. Bisection envelopes of polygons

From Proposition 2.5, we know that the bisection envelope of a bisection convex curve is the midpoint locus of the bisecting chords of its interior area. We apply this fact to the computation of the bisection envelope of a bisection convex polygon.

Let  $A(\theta)$ ,  $B(\theta)$  be the endpoints of the bisecting chord with direction  $\theta$ , with  $A(\theta + \pi) = B(\theta) = A(\theta - \pi)$ . If  $\mathcal{S}$  is a polygon, we can split the interval  $[0, \pi)$  into a finite number of subintervals  $[0, \theta_1)$ ,  $[\theta_1, \theta_2)$ ,  $\dots$ ,  $[\theta_n, \pi)$  such that on each subinterval, the locus of each of  $A(\theta)$  and  $B(\theta)$  is a line segment.

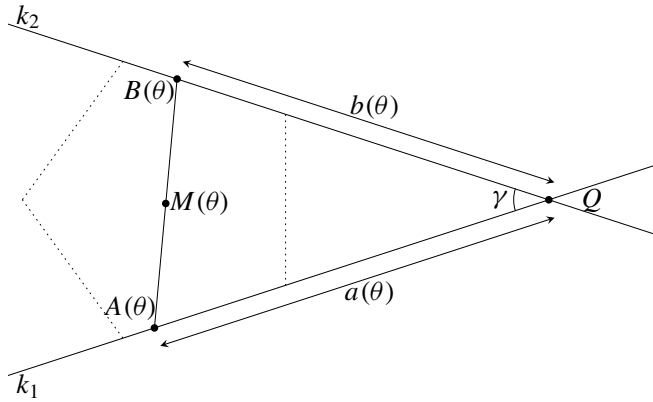
**Proposition 3.1.** *The locus of points  $M(\theta) = (A(\theta) + B(\theta))/2$  over any of the intervals  $[\theta_i, \theta_{i+1})$  is either a section of a hyperbola or a point.*

*Proof.* Let all points  $A(\theta)$  lie on line  $k_1$  and all points  $B(\theta)$  lie on line  $k_2$ . If  $k_1$  and  $k_2$  are parallel, it follows from Corollary 2.7 that the locus of  $M(\theta)$  is a point. Otherwise,  $k_1$  and  $k_2$  meet at a point  $Q$ . Let  $a(\theta) = d(A(\theta), Q)$  and  $b(\theta) = d(B(\theta), Q)$ .

If we construct the triangles  $\triangle A(\theta)QB(\theta)$ , they each have area  $\frac{1}{2}a(\theta)b(\theta) \sin \gamma$ , where  $\gamma$  is the angle between  $k_1$  and  $k_2$ , a constant; see Figure 3. Furthermore, the chords  $\overline{A(\theta)B(\theta)}$  are area preserving on  $\mathcal{S}$ ; therefore, the triangles have constant area, or

$$\frac{1}{2}a(\theta)b(\theta) \sin \gamma = ca(\theta)b(\theta) = \frac{2c}{\sin \gamma} = c', \tag{3-1}$$

for some constant  $c'$ .



**Figure 3.** The situation considered in the proof of Proposition 3.1

Thus there exist distinct unit vectors  $w_1, w_2$  parallel to  $k_1, k_2$  respectively such that

$$M(\theta) = Q + \frac{a(\theta)w_1 + b(\theta)w_2}{2} = Q + \frac{a(\theta)w_1 + (c'/a(\theta))w_2}{2}. \tag{3-2}$$

We see that  $M(\theta)$  is a linear transformation of the set of points

$$\left( a(\theta), \frac{c'}{a(\theta)} \right),$$

which represents a section of a hyperbola. Note that the image of a hyperbola under a linear transformation is itself a hyperbola. □

**Proposition 3.2.** *On any such interval  $[\theta_i, \theta_i + 1)$ , if the locus of  $M(\theta)$  is a section of a hyperbola, the asymptotes of the hyperbola are the two lines  $k_1$  and  $k_2$ , where  $k_1$  and  $k_2$  contain all  $A(\theta)$  and  $B(\theta)$ , respectively.*

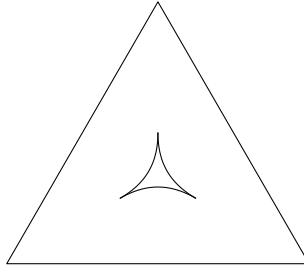
The proof of Proposition 3.2 is left to the reader.

**Proposition 3.3.** *The bisection envelope  $\mathcal{B}$  of a polygon  $S$  is the union of a finite number of sections of hyperbolas. Let the set of all asymptotes of these hyperbolas be  $H$ , and let the set of all lines that contain the sides of  $S$  be  $G$ . Then  $H \subseteq G$ , with equality if no two lines in  $G$  are parallel.*

This follows from the previous two propositions.

This makes the calculation of a bisection envelope of a polygon significantly easier — one must only find the bisecting lines through the vertices and their midpoints; this then strictly defines each of the hyperbolas on each section  $[\theta_i, \theta_{i+1})$ .

**Example 3.4.** The bisection envelope of an equilateral triangle  $\triangle ABC$  of side length two centered on the origin with  $A = (0, 2/\sqrt{3})$ ,  $B = (1, -1/\sqrt{3})$ , and  $C = (-1, -1/\sqrt{3})$  can be found as follows.



**Figure 4.** The bisection envelope of an equilateral triangle found in Example 3.4.

Let  $A', B', C'$  be on the triangle such that the chord  $AA'$  is bisecting, and so forth. The bisection envelope is split into 3 sections: a section of a hyperbola from  $(A + A')/2$  to  $(B + B')/2$  with asymptotes  $AC$  and  $BC$ , and two other congruent hyperbolic sections; see Figure 4.

Specifically, we have

$$A' = \left(0, -\frac{1}{\sqrt{3}}\right), \quad B' = \left(-\frac{1}{2}, \frac{1}{2\sqrt{3}}\right), \quad C' = \left(\frac{1}{2}, \frac{1}{2\sqrt{3}}\right).$$

Therefore

$$\frac{A+A'}{2} = \left(0, \frac{1}{2\sqrt{3}}\right), \quad \frac{B+B'}{2} = \left(\frac{1}{4}, -\frac{1}{4\sqrt{3}}\right), \quad \frac{C+C'}{2} = \left(-\frac{1}{4}, -\frac{1}{4\sqrt{3}}\right). \quad (3-3)$$

The three hyperbolas, from  $A$  to  $B$ ,  $B$  to  $C$ , and  $C$  to  $A$  respectively, are

$$\left(\left(y - \frac{2}{\sqrt{3}}\right) + \sqrt{3}x\right)\left(y + \frac{1}{\sqrt{3}}\right) = c_1, \quad (3-4)$$

$$\left(\left(y - \frac{2}{\sqrt{3}}\right) - \sqrt{3}x\right)\left(\left(y - \frac{2}{\sqrt{3}}\right) + \sqrt{3}x\right) = c_2, \quad (3-5)$$

$$\left(y + \frac{1}{\sqrt{3}}\right)\left(\left(y - \frac{2}{\sqrt{3}}\right) - \sqrt{3}x\right) = c_3. \quad (3-6)$$

By plugging in (3-3) above, we can find

$$c_1 = -\frac{3}{4}, \quad c_2 = \frac{3}{2}, \quad c_3 = -\frac{3}{4}.$$

This defines the bisection envelope fully.

**Theorem 2.** *A polygon with no mutually parallel sides is uniquely defined by its bisection envelope.*

*Proof.* From observations in Proposition 3.1, the assumptions in the theorem give us that the bisection envelope of this polygon does not contain any static points. This is to say, over each of the intervals  $[\theta_i, \theta_{i+1})$ ,  $M(\theta)$  is not a point but a section of a

hyperbola, and therefore, there exists a bijection between the points on the interval  $[\theta_i, \theta_{i+1})$  and the points on the locus of the restriction of  $M(\theta)$  to that range.

From Proposition 3.2, we know the two lines  $k_1, k_2$ , upon which  $A(\theta), B(\theta)$  must lie.  $A(\theta)$  and  $B(\theta)$  must each lie on the line in direction  $\theta$  through  $M(\theta)$ , a line distinct from  $k_1$  and  $k_2$ , so the points  $A(\theta), B(\theta)$  are strictly determined over the interval  $[\theta_i, \theta_{i+1})$ . This can be done for every such interval, and the union of all such intervals is  $[0, \pi)$ ; thus we achieve uniqueness for the loci of  $A(\theta), B(\theta)$  over all  $\theta$ , giving the result.  $\square$

#### 4. Backwards construction

The natural question arises: are there multiple curves with the same bisection envelope? Given a bisection envelope  $\mathcal{B}$ , can we generate all suitable curves with  $\mathcal{B}$  as their bisection envelope?

First we ask, what curves can be bisection envelopes? Suppose that  $\mathcal{B}$  is a bisection envelope associated to some strictly bisection convex curve  $\mathcal{S}$  which is piecewise of class  $C^1$  with a finite number of pieces. Its bisecting lines are  $\mathcal{L} = \{l_\theta\}$ , as explained earlier.

Define  $f : \mathbb{R} \rightarrow \mathbb{R}^2$  by  $f(\theta) = \lim_{\epsilon \rightarrow 0} l_\theta \cap l_{\theta+\epsilon}$ . From Proposition 2.5, we know this is the midpoint of the bisecting chord in direction  $\theta$ , described by the function  $M(\theta)$  presented in Proposition 3.1. Then we have:

**Proposition 4.1.** *The function  $f$  is continuous.*

*Proof.* This follows immediately from the definition  $M(\theta) := (A(\theta) + B(\theta))/2$ , as we have from Proposition 2.4 that  $A(\theta)$  and  $B(\theta)$  vary continuously along  $\mathcal{S}$ .  $\square$

Since  $\mathcal{S}$  has tangents which vary continuously everywhere except a finite number of points, by Proposition 2.6,  $f'$  is defined everywhere but a finite number of points, and where it is defined, it is of the form  $v_\theta(\cos \theta, \sin \theta)$  for a scalar  $v_\theta$ . Therefore, it is possible to define  $f$  as the Lebesgue integral of  $f'$ , giving

$$f(\theta) := f(0) + \int_0^\theta v_t(\cos t, \sin t) dt. \quad (4-1)$$

The value of  $f(0)$  is unimportant — it can just be set to the origin.

Now we generate a curve  $\mathcal{S}'$  from  $f$  and a radius function  $r : \mathbb{R} \rightarrow \mathbb{R}$ , with  $r(\theta + \pi) = r(\theta)$  and  $r(\theta) > 0$ . We define the function  $r$  to be continuous and piecewise of class  $C^1$  with a finite number of pieces.

Define  $\mathcal{S}'$  to be the image of the function

$$g(\theta) := f(\theta) + r(\theta)(\cos \theta, \sin \theta). \quad (4-2)$$

We have then that  $S'$  is continuous, compact, and piecewise of class  $C^1$  with a finite number of pieces; however, we do not have that it is simple. It is clear that

$$g(\theta + 2\pi) = g(\theta) \quad \text{and} \quad \frac{g(\theta) + g(\theta + \pi)}{2} = f(\theta).$$

Thus the chords  $\overline{g(\theta)g(\theta + \pi)}$  are area-preserving if  $S'$  has a well-defined interior and if the chords lie strictly within this interior except at their endpoints, that is, if  $S'$  is simple and bisection convex. The remainder of Section 4 is concerned with the proof of Theorem 3.

**Theorem 3.** *Let  $f, g$  be defined as above.*

*Let  $S'$  be the image of  $g$  and  $\mathcal{B}$  be the image of  $f$ . If  $S' \cap \mathcal{B} = \emptyset$ , then  $\mathcal{B}$  is the bisection envelope of  $S'$ .*

To prove Theorem 3, we use a consequence of the following result.

**Theorem 4.** *Let  $f$  be defined as above with image  $\mathcal{B}$ . Let  $\mathcal{L}$  be the set of lines  $l_\theta$  through  $f(\theta)$  in direction  $\theta$  for all  $\theta$ .*

*Given a point  $P \in \mathbb{R}^2 \setminus \mathcal{B}$ , let  $m_P$  be the number of lines in  $\mathcal{L}$  for which  $P$  lies on  $l_\theta$ , and let  $w(P)$  be the winding number of  $f$  around  $P$  with  $\theta$  increasing over an interval of  $\pi$ . Then*

$$m_P = -2w(P) + 1. \tag{4-3}$$

The proof of Theorem 4 begins by looking at the winding number of a simpler function.

**Lemma 4.2.** *Define the function*

$$f_P(\theta) = (f(\theta) - P) \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}.$$

*If  $f(\theta) \neq P$  for all  $\theta$  then, over the interval  $0 \leq \theta < 2\pi$ , let  $n_P$  be the number of values of  $\theta$  for which  $f_P(\theta)$  lies on the  $x$ -axis, and let  $w_P$  be the winding number of  $f_P(\theta)$  about the origin. Then*

$$w_P = -\frac{1}{2}n_P. \tag{4-4}$$

*Proof.* We have

$$\begin{aligned} f'_P(\theta) &= f'(\theta) \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} + (f(\theta) - P) \begin{pmatrix} -\sin \theta & -\cos \theta \\ \cos \theta & -\sin \theta \end{pmatrix} \\ &= v_\theta(1, 0) + f_P(\theta) \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \\ &= (v_\theta + y, -x), \quad \text{where } f_P(\theta) = (x, y). \end{aligned} \tag{4-5}$$

Note that if  $x > 0$ ,  $y' < 0$ , and vice versa.

Now consider  $f_P(\theta)$  over the half-open interval  $[0, 2\pi)$ . We have  $f_P(\theta + \pi) = f_P(\theta)$ , so the image of  $f_P$  is a closed loop, and  $f_P(\theta)$  is never equal to  $(0, 0)$ , so it has a winding number about the origin.

Let  $\theta_1 < \theta_2 < \dots < \theta_{n_P}$  be the values of  $\theta$  for which  $f_P(\theta)$  lies on the x-axis. Let  $f_P(\theta_1) = (x_1, 0)$  and so on, with  $x_i \neq 0$  by assumption. Then

$$\begin{aligned} x_i &= g - f'_P(\theta_i) \cdot (0, 1) = - \lim_{h \rightarrow 0^+} \frac{f(\theta_i + h) \cdot (0, 1) - f(\theta_i) \cdot (0, 1)}{h} \\ &= g - \lim_{h \rightarrow 0^+} \frac{f(\theta_i + h) \cdot (0, 1)}{h}. \end{aligned}$$

Similarly,

$$x_{i+1} = \lim_{\lambda \rightarrow 0^+} \frac{f(\theta_i - \lambda) \cdot (0, 1)}{\lambda}.$$

But in the domain  $(\theta_i, \theta_{i+1})$  we have that  $f(\theta) \cdot (0, 1)$  is continuous and, by our choices of  $\theta_i$ , nonzero, so it has constant sign. Therefore, for all  $h, \lambda$  sufficiently small and greater than zero,

$$\text{sign}(f(\theta_i + h) \cdot (0, 1)) = \text{sign}(f(\theta_{i+1} - \lambda) \cdot (0, 1)).$$

Thus

$$\text{sign } x_i = - \text{sign} \frac{f(\theta_i + h) \cdot (0, 1)}{h} = - \text{sign} \frac{f(\theta_{i+1} - \lambda) \cdot (0, 1)}{\lambda} = - \text{sign } x_{i+1}.$$

Therefore the  $x_i$  alternate signs. This also implies  $n_P$  is even and  $x_i$  is positive for  $n_P/2$  values.

The winding number of a curve  $\Gamma$  about a point  $P$  can be calculated descriptively by fixing a ray  $R$  from  $P$  in any direction and counting the number of intersections of  $\Gamma$  with  $R$ . For each intersection where the derivative is counterclockwise about  $P$ , we add 1, and where the derivative is clockwise, we subtract 1. The final total is the winding number. Note that if the derivative is along the ray or zero at any intersections, a more subtle approach is required, but this is not the case here.

If we fix the ray from the origin along the x-axis in positive direction for  $f_P$ , we see from (4-5) that at each intersection the derivative is counterclockwise about the origin; therefore  $w_P = -\frac{1}{2}n_P$ .  $\square$

Now we show the relation between the winding numbers of  $f(\theta)$  about  $P$  and  $f_P(\theta)$  about the origin.

**Lemma 4.3.** *Let the winding number of  $f(\theta)$  about  $P$  over the interval  $[0, \pi)$  be  $w(P)$  and the winding number of  $f_P(\theta)$  about the origin over the interval  $[0, 2\pi)$  be  $w_P$ . Then*

$$w_P = 2w(P) - 1. \tag{4-6}$$

*Proof.* An alternative method of determining the winding number of a function relies on the calculation of an integral; several forms exist, although this proof uses the form

$$\frac{1}{2\pi} \int_a^b \frac{f'(x) \cdot \left( (f(x) - P) \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \right)}{|f(x) - P|^2} dx \tag{4-7}$$

for a function  $f(x)$  about  $P$  on the interval  $(a, b)$ . Now we calculate

$$\begin{aligned} w_P &= \frac{1}{2\pi} \int_0^{2\pi} \frac{f'_P(\theta) \cdot \left( (f_P(\theta) - 0) \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \right)}{|f_P(\theta) - 0|^2} d\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} \frac{\left( f'(\theta) \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} + (f(\theta) - P) \begin{pmatrix} -\sin \theta & -\cos \theta \\ \cos \theta & -\sin \theta \end{pmatrix} \right)}{|f(\theta) - P|^2} \cdot \left( (f(\theta) - P) \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \right) d\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} \frac{\left( f'(\theta) \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \right) \cdot \left( (f(\theta) - P) \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \right)}{|f(\theta) - P|^2} d\theta \\ &\quad + \frac{1}{2\pi} \int_0^{2\pi} \frac{\left( (f(\theta) - P) \begin{pmatrix} -\sin \theta & -\cos \theta \\ \cos \theta & -\sin \theta \end{pmatrix} \right) \cdot \left( (f(\theta) - P) \begin{pmatrix} \sin \theta & \cos \theta \\ -\cos \theta & \sin \theta \end{pmatrix} \right)}{|f(\theta) - P|^2} d\theta. \end{aligned}$$

For the first half of this sum we note that  $\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$  and  $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$  commute, and recall that a dot product is unaffected by an isometry applied to both multiplicands. Furthermore, note that  $f$  is periodic in  $\pi$ , so this integral can be split into two identical parts. For the second half of the sum, recall that  $v \cdot (-v) = -|v|^2$ . This allows us to simplify to

$$\begin{aligned} w_P &= 2 \left( \frac{1}{2\pi} \int_0^\pi \frac{f'(\theta) \cdot \left( (f(\theta) - P) \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \right)}{|f(\theta) - P|^2} d\theta \right) + \frac{1}{2\pi} \int_0^{2\pi} -1 d\theta \\ &= 2w(P) - 1. \end{aligned} \quad \square$$

The results of the two preceding lemmas can be combined to achieve Theorem 4.

*Proof of Theorem 4.* When  $P$  is on  $l_\theta$ ,  $f_P(\theta)$  lies on the x-axis, but

$$f_P(\theta + \pi) = -f_P(\theta) \neq (0, 0) \quad \text{and} \quad l_{\theta+\pi} = l_\theta,$$

so the number  $m_P$  of distinct lines  $l_\theta$  containing  $P$  is equal to half the number of times  $f_P(\theta)$  lies on the x-axis in the interval  $[0, 2\pi)$ . Using Lemmas 4.2 and 4.3,

$$m_P = n_P/2 = -w_P = -2w(P) + 1. \quad \square$$

**Corollary 4.4.** *Every point  $P$  in the exterior of  $\mathcal{B}$  lies on precisely one bisecting line  $l_\theta$ .*

*Proof.* Since  $P$  is on the exterior of  $\mathcal{B}$ , we have  $w(P) = 0$ , and the result follows from Theorem 4.  $\square$

**Remark 4.5.** This also implies that no bisection envelope can have strictly positive winding number about any point, or the value  $m_P$  would be negative and have no meaning. Intuitively, this could be observed from  $f_P$ , which may not wind counterclockwise about the origin.

Theorem 4 can be used to show the first step in proving Theorem 3.

**Lemma 4.6.** *If  $S'$  lies on the exterior of  $\mathcal{B}$ , then it is not self-intersecting and each  $l_\theta$  intersects  $S'$  exactly twice, at  $g(\theta)$  and  $g(\theta + \pi)$ .*

*Proof.* If either of these conditions are false, there exist two lines  $l_{\theta_1}, l_{\theta_2}$  that intersect at some point on  $S'$ , say at  $P$ . But  $S'$ , and thus  $P$ , lies on the exterior of  $\mathcal{B}$ ; thus  $w(P) = 0$ . By Theorem 4, this leads to the contradiction

$$2 \leq m_P = 2(-0) + 1 = 1. \quad \square$$

**Lemma 4.7.** *Given two continuous, compact curves  $C_1, C_2 \in \mathbb{R}^2$ , if  $C_2$  lies fully in the interior of  $C_1$ , then for each  $P \in \mathbb{R}^2$ , there exists a point  $P_1 \in C_1$  such that for all  $P_2 \in C_2$ ,*

$$d(P_1, P) > d(P_2, P).$$

*Proof.* If  $P_2$  lies on the interior of  $C_1$ , then there is a ball around  $P_2$  that lies on the interior of  $C_1$ . The ray starting at  $P$  passing through  $P_2$  extends to points past  $P_2$  but still in the interior of  $C_1$ . Since  $C_1$  is bounded, eventually this ray must intersect  $C_1$  at a point  $Q$ , and  $d(Q, P) > d(P_2, P)$ .

Let  $P_1$  be a point on  $C_1$  such that  $d(P_1, P)$  is maximal (this can be done as  $C_1$  is compact). Then

$$d(P_1, P) \geq d(Q, P) > d(P_2, P)$$

for all  $P_2$ . This can be done for every point  $P$ .  $\square$

**Lemma 4.8.**  *$S'$  cannot lie fully in the interior of  $\mathcal{B}$ .*

*Proof.* From the definition of  $g$ , for a point  $P_1$  on  $\mathcal{B}$ , there exist points  $P_2 = P_1 + a$  and  $P'_2 = P_1 - a$  on  $S'$  for some nonzero vector  $a$  ( $r$  is defined to be greater than zero); then  $P_2, P_1$ , and  $P'_2$  are collinear in that order.

It follows that given any reference point  $P$ ,  $P_1$  cannot be the furthest of these points from  $P$ ; thus by the contrapositive of Lemma 4.7,  $S'$  is not fully in the exterior of  $\mathcal{B}$ .  $\square$



*Proof of Theorem 3.* If  $S' \cap \mathcal{B} = \emptyset$ , then  $S'$  lies fully in the exterior of  $\mathcal{B}$ —it cannot lie in the interior by Lemma 4.8. By Lemma 4.6,  $S'$  must not be self-intersecting, so it has a well-defined interior, and each line  $l_\theta$  touches  $S'$  at exactly two points. Thus the chords  $\overline{g_\theta g_{\theta+\pi}}$  are fully contained in the interior of  $S'$ . By Proposition 2.5, they are area preserving, and  $\overline{g_\theta g_{\theta+\pi}} = \overline{g_{\theta+\pi} g_{\theta+2\pi}}$ , so they are bisecting lines of the interior of  $S'$ . From the definitions,  $f(\theta)$  is the midpoint of  $g(\theta)$  and  $g(\theta + \pi)$ , so again by Proposition 2.5,  $\mathcal{B}$  is the bisection envelope of  $S'$ .  $\square$

**Remark 4.9.** Note that  $\mathcal{S}$  is strictly bisection convex; therefore by Proposition 2.6, there are no points on  $\mathcal{B}$  where the limit of  $|f'(\theta)|$  is infinite. However,  $\mathcal{B}$  is also the bisection envelope of  $S'$ , so  $S'$  is also strictly bisection convex.

**Remark 4.10.** If the radius function  $r$  is sufficiently large,  $S'$  cannot intersect  $\mathcal{B}$ . This implies that for any strictly bisection convex  $\mathcal{S}$ , there are an infinite number of other strictly bisection convex curves  $S'$  that share its bisection envelope, each generated by a different  $r$ .

### 5. Relations between areas

Using the construction from the previous section, we now determine the interior area of  $S'$  as the sum of two integrals, one involving  $r(\theta)$  and another that gives the interior area of  $f(\theta)$ . Note that we assume  $f$  and  $g$  are differentiable almost everywhere throughout this section.

We define (and denote) interior area of a closed, continuous curve purely based upon the line integral

$$\mathcal{A}(\Gamma) = \frac{1}{2} \oint_{\Gamma} x \, dy - y \, dx \tag{5-1}$$

irrespective of whether the curve has a well-defined interior. Note that whenever the curve  $\Gamma$  is simple, that is, when discussion of area makes sense, this area function gives its exact area, positive or negative depending on the direction we integrate about  $\Gamma$ . Also note that this integral functions equivalently to the double integral

$$\iint_{\mathbb{R}^2 \setminus \Gamma} w(\Gamma, P) \, dx \, dy, \tag{5-2}$$

where  $P = (x, y)$  and  $w(\Gamma, P)$  is the winding number of  $\Gamma$  about  $P$ .

**Theorem 5.** 
$$\mathcal{A}(S') = \int_0^{2\pi} \frac{r^2(\theta)}{2} \, d\theta + 2\mathcal{A}(\mathcal{B}). \tag{5-3}$$

*Proof.* We recall that  $S'$  is parametrized by

$$g(\theta) = f(0) + \int_0^\theta v_t(\cos t, \sin t) \, dt + r(\theta)(\cos \theta, \sin \theta).$$

Since  $f(0)$  is arbitrary, we take it to be zero.

Next we take the derivative and separate the  $x$  and  $y$  components, giving

$$g'(\theta) = (v_\theta \cos \theta + r'(\theta) \cos \theta - r(\theta) \sin \theta, v_\theta \sin \theta + r'(\theta) \sin \theta + r(\theta) \cos \theta).$$

We expand and simplify  $\mathcal{A}(\mathcal{S}')$  using standard trigonometric identities.

$$\begin{aligned} \mathcal{A}(\mathcal{S}') &= \frac{1}{2} \oint_{\mathcal{S}} x dy - y dx = \frac{1}{2} \int_0^{2\pi} \left( x \frac{dy}{d\theta} - y \frac{dx}{d\theta} \right) d\theta \\ &= \frac{1}{2} \int_0^{2\pi} \left( \left( \int_0^\theta v_t \cos t dt + r(\theta) \cos \theta \right) (v_\theta \sin \theta + r'(\theta) \sin \theta + r(\theta) \cos \theta) \right. \\ &\quad \left. - \left( \int_0^\theta v_t \sin t dt + r(\theta) \sin \theta \right) (v_\theta \cos \theta + r'(\theta) \cos \theta - r(\theta) \sin \theta) \right) d\theta \\ &= \int_0^{2\pi} \frac{r^2(\theta)}{2} d\theta + \frac{1}{2} \int_0^{2\pi} \int_0^\theta v_t v_\theta (\sin(\theta - t)) dt d\theta \\ &\quad + \frac{1}{2} \int_0^{2\pi} \left( r'(\theta) \int_0^\theta v_t \sin(\theta - t) dt + r(\theta) \int_0^\theta v_t \cos(\theta - t) dt \right) d\theta. \end{aligned} \quad (5-4)$$

Observe that, from the points in Definition 2.10,

$$\begin{aligned} \int_0^{\theta+\pi} v_t \sin((\theta + \pi) - t) dt &= \int_\pi^{\theta+\pi} -v_{t+\pi} \sin((\theta + \pi) - t) dt \\ &= - \int_0^\theta v_t \sin(\theta - t) dt. \end{aligned} \quad (5-5)$$

Similarly,

$$\int_0^{\theta+\pi} v_t \cos((\theta + \pi) - t) dt = - \int_0^\theta v_t \cos(\theta - t) dt. \quad (5-6)$$

By splitting the integrals and replacing variables, the final line of (5-4) can be rewritten to give

$$\begin{aligned} &\frac{1}{2} \int_0^\pi r'(\theta) \int_0^\theta v_t \sin(\theta - t) dt d\theta + \frac{1}{2} \int_0^\pi r'(\theta + \pi) \int_0^{\theta+\pi} v_t \sin(\theta + \pi - t) dt d\theta \\ &+ \frac{1}{2} \int_0^\pi r(\theta) \int_0^\theta v_t \cos(\theta - t) dt d\theta + \frac{1}{2} \int_0^\pi r(\theta + \pi) \int_0^{\theta+\pi} v_t \cos(\theta + \pi - t) dt d\theta. \end{aligned}$$

As  $r(\theta + \pi) = r(\theta)$ , this can further be written as

$$\begin{aligned} &\frac{1}{2} \int_0^\pi r'(\theta) \left( \int_0^\theta v_t \sin(\theta - t) dt + \int_0^{\theta+\pi} v_t \sin(\theta + \pi - t) dt \right) d\theta \\ &\quad + \frac{1}{2} \int_0^\pi r(\theta) \left( \int_0^\theta v_t \cos(\theta - t) dt + \int_0^{\theta+\pi} v_t \cos(\theta + \pi - t) dt \right) d\theta. \end{aligned}$$

However, from (5-5) and (5-6) this entire expression amounts to zero. From (5-4), we are left with

$$\mathcal{A}(S') = \int_0^{2\pi} \frac{r^2(\theta)}{2} d\theta + \frac{1}{2} \int_0^{2\pi} \int_0^\theta v_t v_\theta (\sin(\theta - t)) dt d\theta. \tag{5-7}$$

In a similar fashion to the above, the second term can be rewritten as

$$\frac{1}{2} \int_0^\pi v_\theta \left( \int_0^\theta v_t \sin(\theta - t) dt - \int_0^{\theta+\pi} v_t \sin(\theta + \pi - t) dt \right) d\theta.$$

Note the change in the negative sign, as  $v_{\theta+\pi} = -v_\theta$ . By (5-5) this is equal to

$$2 \left( \frac{1}{2} \int_0^\pi \int_0^\theta v_t v_\theta \sin(\theta - t) dt d\theta \right). \tag{5-8}$$

Now applying (5-1) to  $\mathcal{B}$ , we recall that  $\mathcal{B}$  is parametrized by

$$f(\theta) = f(0) + \int_0^\theta v_t (\cos t, \sin t) dt,$$

with derivative

$$f'(\theta) = (v_\theta \cos \theta, v_\theta \sin \theta).$$

Thus

$$\begin{aligned} \mathcal{A}(\mathcal{B}) &= \frac{1}{2} \oint_{\mathcal{B}} x dy - y dx = \frac{1}{2} \int_0^\pi \left( x \frac{dy}{d\theta} - y \frac{dx}{d\theta} \right) d\theta \\ &= \frac{1}{2} \int_0^\pi \left( \int_0^\theta v_t \cos t v_\theta \sin \theta dt - \int_0^\theta v_t \sin t v_\theta \cos \theta dt \right) d\theta \\ &= \frac{1}{2} \int_0^\pi \int_0^\theta v_t v_\theta \sin(\theta - t) dt d\theta. \end{aligned} \tag{5-9}$$

Combining (5-7), (5-8), and (5-9), it is finally achieved that

$$\mathcal{A}(S') = \int_0^{2\pi} \frac{r^2(\theta)}{2} d\theta + 2\mathcal{A}(\mathcal{B}). \quad \square$$

This formula may be useful in determining the area of a bisection envelope where the integral (5-9) is much more difficult than finding  $r(\theta)$  then calculating (5-3).

A property of  $\mathcal{B}$  described in Remark 4.5 allows us to bound  $\mathcal{A}(\mathcal{B})$ .

**Proposition 5.1.**  $\mathcal{A}(\mathcal{B}) \leq 0.$

*Proof.* Remark 4.5 notes that, for all  $P \notin \mathcal{B}$ ,

$$w(\mathcal{B}, P) \leq 0.$$

Thus from (5-2),

$$\mathcal{A}(\mathcal{B}) = \iint_{\mathbb{R}^2 \setminus \mathcal{B}} w(\mathcal{B}, P) \, dx \, dy \leq 0. \quad \square$$

**Proposition 5.2.** *Let  $S'$  be piecewise of class  $C^1$  with a finite number of pieces. If  $\mathcal{A}(\mathcal{B}) = 0$ , then  $\mathcal{B}$  is a point.*

*Proof.* If  $\mathcal{A}(\mathcal{B}) = 0$ , then from the reasoning in Proposition 5.1,  $w(\mathcal{B}, P) = 0$  for all  $P$  not on  $\mathcal{B}$ .

Consider three bisecting lines  $l_{\theta_1}, l_{\theta_2}, l_{\theta_3}$  with mutual intersections  $A, B, C$ . Assume the three points are distinct. From continuity, we have that all points  $P$  in the interior of  $\triangle ABC$  lie on at least three lines  $l_{\theta}$ . By Theorem 3, this implies that for all such  $P$ ,  $w(P) \leq -1$ , and therefore  $P$  must be on  $\mathcal{B}$ . Hence,  $\mathcal{B}$  is a space-filling curve on some subset of  $\mathbb{R}^2$  that contains  $\triangle ABC$ . However,  $f$  is of class  $C^1$  at all but a finite number of points, so it cannot be a space-filling curve.

It follows that any three bisecting lines are concurrent, and thus, all bisecting lines are concurrent, and  $\mathcal{B}$  is a point. □

**Corollary 5.3.** *Of all bisection convex curves  $S'$  piecewise of class  $C^1$  with a finite number of pieces such that*

$$\int_0^{2\pi} \frac{r^2(\theta)}{2} \, d\theta = k$$

*for some fixed  $k$ , those with maximal interior area have  $180^\circ$  rotational symmetry.*

*Proof.* From Theorem 5 and Proposition 5.1, these curves clearly have maximal interior area when  $\mathcal{A}(\mathcal{B}) = 0$ . By Proposition 5.2, this is only possible if  $\mathcal{B}$  is a point, say  $P$ . From the definition of  $g$ ,  $S'$  has  $180^\circ$  rotational symmetry about  $P$ . □

**Remark 5.4.** The proof of Corollary 5.3 shows that if we drop the restriction that  $S'$  is piecewise of class  $C^1$  with a finite number of pieces and rather assume it is only piecewise of class  $C^1$ , then the bisection envelope consists of all the points of intersection between bisecting lines and this envelope might be space-filling. We are unable to rule out the possibility of a space-filling bisection envelope and leave it as an open question: can  $f$  be differentiable almost everywhere and space-filling?

Lastly, we use Theorem 5 to find the internal area of the bisection envelope of an equilateral triangle calculated in Example 3.4.

**Example 5.5.** The bisection envelope of a triangle is not self-intersecting; therefore its interior area is well-defined and is recognized to be  $-\mathcal{A}(\mathcal{B})$ . Rearranging Theorem 5, we have

$$-\mathcal{A}(\mathcal{B}) = \frac{\int_0^{2\pi} r^2(\theta)/2 \, d\theta - \mathcal{A}(S')}{2}.$$

Now  $\mathcal{A}(S')$  is the area of an equilateral triangle with side length 2 or  $\sqrt{3}$ . Also, by symmetry,  $r$  has period  $\pi/3$ , and therefore we rewrite

$$-\mathcal{A}(B) = 3 \int_0^{\pi/3} \frac{r^2(\theta)}{2} d\theta - \frac{\sqrt{3}}{2}. \tag{5-10}$$

Rotation of the triangle has no effect on area, and thus we rotate so that the three medians have directions  $0, \pi/3, 2\pi/3$  with  $A, B, C$  being the vertices that lie on the respective medians.

Let  $A(\theta), B(\theta)$  be the intersection points of  $l_\theta$  with the triangle, where  $A(0) = A, B(\pi/3) = B$ . Let  $a(\theta) = d(A(\theta), C)$  and  $b(\theta) = d(B(\theta), C)$ . Since the  $A(\theta)B(\theta)$  are bisecting chords, we have  $\frac{1}{2}a(\theta)b(\theta) \sin(\pi/3) = \sqrt{3}/2$ , which implies

$$a(\theta)b(\theta) = 2. \tag{5-11}$$

We now apply the sine and cosine laws to get  $2r(\theta) \sin\left(\frac{\pi}{2} - \theta\right) = a(\theta) \sin \frac{\pi}{3}$  on the one hand, which yields

$$a(\theta) = \frac{4}{\sqrt{3}}r(\theta) \cos \theta, \tag{5-12}$$

and on the other hand

$$4r^2(\theta) = a^2(\theta) + b^2(\theta) - 2a(\theta)b(\theta) \cos \frac{\pi}{3}. \tag{5-13}$$

Combining (5-11), (5-12), and (5-13) we have

$$\left(2 - \frac{8}{3} \cos^2 \theta\right)r^4(\theta) + r^2(\theta) - \frac{3}{8 \cos^2 \theta} = 0. \tag{5-14}$$

Thus we find

$$\frac{r^2(\theta)}{2} = \frac{1 \pm \sqrt{3} \tan \theta}{\frac{32}{3} \cos^2 \theta - 8}. \tag{5-15}$$

We choose the  $\pm$  to be a  $-$ , otherwise as  $\theta \rightarrow \pi/6$ , we have that  $r^2(\theta)$  goes to infinity. This function is integrable by standard methods by a change of variable to  $u = \cot \theta$  and then through use of partial fractions. We calculate

$$\int_0^{\pi/3} \frac{r^2(\theta)}{2} d\theta = \frac{1}{8} \sqrt{3} \ln(1 + \sqrt{3} \tan \theta) \Big|_0^{\pi/3} = \frac{\sqrt{3}}{4} \ln 2. \tag{5-16}$$

This can now be inserted back into (5-10), giving the result

$$-\mathcal{A}(B) = \frac{3\sqrt{3}}{4} \ln 2 - \frac{\sqrt{3}}{2} \approx 0.03440. \tag{5-17}$$

**Remark 5.6.** As ratios of areas and ratios of lengths along a line are unaffected by linear transformations, the bisection envelope of a curve will remain unchanged under a linear transformation. As any triangle is the image of any other triangle

under some linear transformation, it follows that the ratio  $\mathcal{A}(\mathcal{B}) : \mathcal{A}(S')$  is a constant when  $S'$  is a triangle. Therefore, for all triangles  $S'$ ,

$$\frac{\mathcal{A}(\mathcal{B})}{\mathcal{A}(S')} = \frac{3}{4} \ln 2 - \frac{1}{2} \approx 0.01986. \quad (5-18)$$

In other words, every triangle has a bisection envelope with area roughly a fiftieth of its area.

### Acknowledgements

I would like to thank C. Kenneth Fan for all of the time and effort he has put in making this paper a possibility. From offering initial ideas and working through proofs, to reviewing and editing a final product, he has been a valuable collaborator and mentor. I would also like to thank the referee their insightful comments.

### References

[Fusco and Pratelli 2011] N. Fusco and A. Pratelli, “On a conjecture by Auerbach”, *J. Eur. Math. Soc. (JEMS)* **13**:6 (2011), 1633–1676. MR 2012h:52021 Zbl 1227.49047

Received: 2013-07-30    Revised: 2013-10-23    Accepted: 2013-11-12

noahfp@gmail.com

*Harvard University, 1405 Harvard Yard Mail Center,  
Cambridge, MA 02138, United States*

# Degree 14 2-adic fields

Chad Awtrey, Nicole Miles, Jonathan Milstead,  
Christopher Shill and Erin Strosnider

(Communicated by Nigel Boston)

We study the 590 nonisomorphic degree 14 extensions of the 2-adic numbers by computing defining polynomials for each extension as well as basic invariant data for each polynomial, including the ramification index, residue degree, discriminant exponent, and Galois group. Our study of the Galois groups of these extensions shows that only 10 of the 63 transitive subgroups of  $S_{14}$  occur as a Galois group. We end by describing our implementation for computing Galois groups in this setting, which is of interest since it uses subfield information, the discriminant, and only one other resolvent polynomial.

## 1. Introduction

Hensel's  $p$ -adic numbers are a foundational tool in 21st century number theory, with applications to such areas as number fields, elliptic curves, and representation theory (among others). They are also the subject of much current research themselves, with several studies aimed at classifying arithmetic invariants of finite extensions of the  $p$ -adic numbers. Among the most useful invariants to identify are the ramification index, residue degree, discriminant, and Galois group (of the normal closure) of each extension. For such a pursuit, we can take the following classical result as motivation [Lang 1994, p. 54].

**Theorem 1.1.** *For a fixed prime number  $p$  and positive integer  $n$ , there are only finitely many nonisomorphic extensions of the  $p$ -adic numbers of degree  $n$ .*

When  $p \nmid n$ , all extensions are tamely ramified and are well understood [Jones and Roberts 2006]. Likewise, when  $p = n$ , the situation has been solved since the early 1970s [Amano 1971; Jones and Roberts 2006]. The difficult cases where  $p \mid n$  and  $n$  is composite have been dealt with on a case-by-case basis for low degrees  $n$  and small primes  $p$ . Jones and Roberts [2004; 2006; 2008] have classified the cases where  $n \leq 10$ , and the case of degree 12 is dealt with in [Awtrey 2012; Awtrey and Shill 2013; Awtrey et al.  $\geq 2015a$ ;  $\geq 2015b$ ].

---

MSC2010: 11S15, 11S20.

Keywords: 2-adic, extension fields, Galois group, local field.

In this paper, we are concerned with classifying degree 14 extensions of the 2-adic numbers. In particular, we focus on computing defining polynomials for each field as well as the Galois group for each of these polynomials. The other invariants are straightforward to compute using basic number field commands in [PARI 2012]. In Section 2, we lay the theoretical groundwork for computing Galois groups of  $p$ -adic fields using the theory of ramification groups. A consequence of this section is that every degree 14 extension of  $\mathbb{Q}_2$  has a unique septic subfield. In Section 3, we use the result of Section 2 to compute defining polynomials. In the final section, we discuss our method of determining the Galois groups of the polynomials found in Section 3.

## 2. Ramification groups

The aim of this section is to show that every degree 14 2-adic field has a unique septic subfield. To accomplish this, we introduce the basic properties of ramification groups and use those properties to deduce structural information about degree 14 extensions of  $\mathbb{Q}_2$ . For a more detailed exposition of ramification group theory, see [Serre 1979].

**Definition 2.1.** Let  $L/\mathbb{Q}_p$  be a Galois extension with Galois group  $G$ . Let  $v$  be the discrete valuation on  $L$  and let  $\mathbb{Z}_L$  denote the corresponding discrete valuation ring. For an integer  $i \geq -1$ , we define the  $i$ -th ramification group of  $G$  to be the set

$$G_i = \{\sigma \in G : v(\sigma(x) - x) \geq i + 1 \text{ for all } x \in \mathbb{Z}_L\}.$$

The ramification groups define a sequence of decreasing normal subgroups which are eventually trivial and which give structural information about the Galois group of a  $p$ -adic field. For example, the following result is useful for determining possible Galois groups of  $p$ -adic fields. A proof can be found in [Serre 1979, Chapter 4].

**Lemma 2.2.** Let  $L/\mathbb{Q}_p$  be a Galois extension with Galois group  $G$ , and let  $G_i$  denote the  $i$ -th ramification group. Let  $\mathfrak{p}$  denote the unique maximal ideal of  $\mathbb{Z}_L$  and  $U_0$  the units in  $L$ . For  $i \geq 1$ , let  $U_i = 1 + \mathfrak{p}^i$ .

- (a) For  $i \geq 0$ ,  $G_i/G_{i+1}$  is isomorphic to a subgroup of  $U_i/U_{i+1}$ .
- (b) The group  $G_0/G_1$  is cyclic and isomorphic to a subgroup of the group of roots of unity in the residue field of  $L$ . Its order is prime to  $p$ .
- (c) The quotients  $G_i/G_{i+1}$  for  $i \geq 1$  are abelian groups and are direct products of cyclic groups of order  $p$ . The group  $G_1$  is a  $p$ -group.
- (d) The group  $G_0$  is the semidirect product of a cyclic group of order prime to  $p$  with a normal subgroup whose order is a power of  $p$ .
- (e) The groups  $G_0$  and  $G$  are both solvable.

Suppose  $f$  is an irreducible polynomial of degree 14 defined over  $\mathbb{Q}_2$  and let  $G$  be its Galois group. From Lemma 2.2, we see that  $G$  is a solvable transitive



subgroup of  $S_{14}$ . Furthermore,  $G$  contains a solvable normal subgroup  $G_0$  such that  $G/G_0$  is cyclic. The group  $G_0$  contains a normal subgroup  $G_1$  such that  $G_1$  is a 2-group (possibly trivial). Moreover,  $G_0/G_1$  is cyclic of order dividing  $2^{\lfloor G:G_0 \rfloor} - 1$ . Direct computation on the 63 transitive subgroups of  $S_{14}$  (using [GAP 2008], for example) shows that only 15 of the 63 are possibilities for the Galois group of  $f$ . Using the transitive group notation in [GAP 2008], these 15 groups are  $\text{TransitiveGroup}(14, n)$ , where  $n$  is one of the following possibilities:

$$\{1, 4, 5, 6, 7, 9, 11, 18, 21, 29, 35, 40, 41, 44, 48\}.$$

Showing that every degree 14 extension of  $\mathbb{Q}_2$  has a unique septic subfield amounts to showing that each of the above 15 groups possesses the corresponding group-theoretic property. In particular, let  $K/\mathbb{Q}_2$  be a degree 14 extension defined by an irreducible polynomial  $f$ , and consider the subfields of  $K$  up to isomorphism. The list of the Galois groups of the Galois closures of the proper nontrivial subfields of  $K$  is important for our work. We call this the *subfield Galois group* content of  $K$ , and we denote it by  $\text{sgg}(K)$ .

The  $\text{sgg}$  content of an extension is an invariant of its Galois group. Indeed, suppose the normal closure of  $K/\mathbb{Q}_2$  has Galois group  $G$  and let  $E$  be the subgroup fixing  $K$ . By Galois theory, the nonisomorphic subfields of  $K$  correspond to the intermediate subgroups  $F$ , up to conjugation, such that  $E \leq F \leq G$ . Specifically, if  $K'$  is a subfield and  $F$  is its corresponding intermediate group, then the Galois group of the normal closure of  $K'$  is equal to the permutation representation of  $G$  acting on the cosets of  $F$  in  $G$ . Consequently, it makes sense to speak of the  $\text{sgg}$  content of a transitive subgroup as well.

For each of these 15 groups, we used [GAP 2008] to compute their  $\text{sgg}$  content. We found that 5 of these groups — 4, 7, 40, 41, 48 — had 7T4 in their  $\text{sgg}$  content. This means that polynomials whose Galois group is one of these 5 possibilities must define an extension with a septic subfield whose normal closure has Galois group 7T4. But as we will see in the next section, the only possible Galois groups of degree 7 polynomials over  $\mathbb{Q}_2$  are either 7T1 or 7T3. This means that these 5 groups cannot occur as the Galois group of a degree 14 2-adic field.

Therefore, there are only 10 possible Galois groups of degree 14 extensions of  $\mathbb{Q}_2$ . For each of these possible Galois groups, Table 3 shows their respective  $\text{sgg}$  contents. Notice that each group has exactly one entry of the form 7Tj. This shows that degree 14 extensions of  $\mathbb{Q}_2$  have a unique septic subfield.

### 3. Defining polynomials

As a consequence of Section 2, every degree 14 extension of  $\mathbb{Q}_2$  can be realized uniquely as a quadratic extension of a septic 2-adic field. Defining polynomials for degree 14 2-adic fields are therefore straightforward to compute.

$e$	$G$	poly
1	7T1	$u7 = x^7 - x + 1$
7	7T3	$t7 = x^7 - 2$

**Table 1.** Septic extensions of  $\mathbb{Q}_2$ , including the ramification index  $e$  and Galois group  $G$  of a defining polynomial poly.

First, we compute all septic 2-adic fields. Such fields are tamely ramified and are therefore easy to classify using [Jones and Roberts 2006]. Table 1 shows that there are two septic 2-adic fields, the unramified extension (with cyclic Galois group) and a totally ramified extension (with  $7T3 = C_7 : C_3$  as its Galois group). Next, for each septic 2-adic field, we compute all of its quadratic extensions using [Awtrey 2010]. In each case, there are 511 such quadratic extensions. But some of these 1022 extensions are isomorphic. Using Panayi’s algorithm [Pauli and Roblot 2001], we discard isomorphic extensions to find a total of 590 nonisomorphic degree 14 extensions of  $\mathbb{Q}_2$ . Polynomials are available on request by emailing the first author.

Table 2 contains numerical data on the numbers of these extensions, excluding the unramified extensions of the two septic 2-adic fields. The “base” column references the two polynomials in Table 1. The column  $c$  is the discriminant exponent,  $G$  is the Galois group of the defining polynomial, and  $\#\mathbb{Q}_2^{14}$  is the number of nonisomorphic extensions over  $\mathbb{Q}_2$ . Notice that there are 78 extensions that are ramified quadratic extensions of the unramified septic 2-adic field. There are 510 ramified quadratic extensions of the unique totally ramified septic 7-adic field. These 588 extensions plus the unramified extensions of the two septic 2-adic fields give 590 total degree 14 extensions of  $\mathbb{Q}_2$ . Krasner’s mass formula [1966] verifies that these are all such extensions. We note that the number of extensions can also be verified using an implementation of [Pauli and Roblot 2001] in [PARI 2012].

#### 4. Galois groups

It remains to identify the Galois group over  $\mathbb{Q}_2$  for each of the 590 polynomials. We follow the standard approach for determining Galois groups [Hulpke 1999]. We compute enough group-theoretic and field-theoretic invariants so as to uniquely identify a polynomial with its corresponding Galois group. Our strategy is to divide the above list of 10 groups into smaller pieces that are easily distinguished from each other. Our first division will be at the level of centralizer order. The order of the centralizer in  $S_{14}$  of the Galois group is useful as it corresponds to the size of the automorphism group of the stem field defined by the polynomial. We divide these smaller sets even further based on their sgg content and their parity. The parity of a group  $G$  is  $+1$  if  $G \subseteq A_{14}$  and  $-1$  otherwise. Likewise, the parity

base	$c$	$G$	$\#\mathbb{Q}_2^{14}$	base	$c$	$G$	$\#\mathbb{Q}_2^{14}$
$u7$	14	14T1	2	$t7$	20	14T5	2
$u7$	14	14T6	2	$t7$	20	14T18	8
$u7$	14	14T9	6	$t7$	20	14T44	6
$u7$	14	14T21	7				
$u7$	14	14T29	21	$t7$	22	14T11	2
				$t7$	22	14T18	6
$u7$	21	14T1	4	$t7$	22	14T35	6
$u7$	21	14T9	8	$t7$	22	14T44	18
$u7$	21	14T29	28				
				$t7$	24	14T11	4
$t7$	14	14T11	1	$t7$	24	14T18	12
$t7$	14	14T18	1	$t7$	24	14T35	12
				$t7$	24	14T44	36
$t7$	16	14T11	1				
$t7$	16	14T18	1	$t7$	26	14T11	4
$t7$	16	14T35	1	$t7$	26	14T18	12
$t7$	16	14T44	1	$t7$	26	14T35	28
				$t7$	26	14T44	84
$t7$	18	14T11	2				
$t7$	18	14T18	2	$t7$	27	14T5	4
$t7$	18	14T35	2	$t7$	27	14T18	56
$t7$	18	14T44	2	$t7$	27	14T44	196

**Table 2.** Ramified quadratic extensions of septic 2-adic fields.

of a polynomial  $f$  is  $+1$  if its discriminant is a square in  $\mathbb{Q}_2$  and  $-1$  otherwise. When this information is not enough, we introduce a single resolvent polynomial [Stauduhar 1973] and use information about its irreducible factors over  $\mathbb{Q}_2$ . This resolvent, denoted as  $f_{364}$ , has degree 364. It corresponds to the subgroup  $S_{11} \times S_3$  of  $S_{14}$  and can be computed as a linear resolvent on 3-sets [Soicher and McKay 1985], i.e., as a resultant. It can also be computed in the following way. Let  $f(x)$  define a degree 14 extension over  $\mathbb{Q}_2$ , and let  $r_1, r_2, \dots, r_{14}$  be the roots of  $f$ . Then,

$$f_{364}(x) = \prod_{i=1}^{12} \prod_{j=i+1}^{13} \prod_{k=j+1}^{14} (x - r_i - r_j - r_k).$$

We note that in our search for suitable resolvent polynomials, we also looked at a lower degree linear resolvent (corresponding to the group  $S_2 \times S_{12}$ ), subfields of the field defined by this lower degree resolvent, and other subfield information of  $f_{364}$ . In order to keep the computational difficulty of our algorithm as low as possible, we focused on subfields of degree less than 12, with a preference toward quadratic subfields of the fields defined by the irreducible factors of the linear resolvents.

$G$	parity	$ C_{S_{14}}(G) $	sgg	$f_{364}$	quad subs	$\#\mathbb{Q}_2^{14}$
14T1	-1	14	2T1, 7T1			7
14T5	-1	2	2T1, 7T3			7
14T6	+1	2	7T1	$14^6, 28^2, 56^4$		2
14T21	+1	2	7T1	$14^6, 56^5$		7
14T9	-1	2	7T1	$14^6, 56^5$	one	14
14T29	-1	2	7T1	$14^6, 56^5$	none	49
14T11	+1	2	7T3	$28^2, 42^2, 56, 168$		14
14T35	+1	2	7T3	$42^2, 56^2, 168$		49
14T18	-1	2	7T3	$42^2, 56^2, 168$	one	98
14T44	-1	2	7T3	$42^2, 56^2, 168$	none	343

**Table 3.** Invariant data for possible Galois groups of degree 14 2-adic fields.

Under these constraints, we found the degree 56 factors of  $f_{364}$  to be the smallest degree factors that accomplished our needs.

Table 3 contains all pertinent invariant data for each Galois group. Notice that all groups can be distinguished using parity, centralizer order, sgg content, and the degrees of the factors of  $f_{364}$  except for two sets: 14T9/14T29 and 14T18/14T44. But in both cases, the groups can be distinguished by counting quadratic subfields of the fields defined by the degree 56 factors of  $f_{364}$ . In these two cases, we have also verified Galois group computations with [Milstead et al. 2015] by computing sizes of splitting fields. As before, we include the column  $\#\mathbb{Q}_2^{14}$ , which represents the number of nonisomorphic extensions over  $\mathbb{Q}_2$  with the corresponding Galois group (which can also be inferred from Table 2). The other columns are defined as follows:  $|C_{S_{14}}(G)|$  gives the size of the centralizer of the group in  $S_{14}$ , sgg gives the sgg content of the group,  $f_{364}$  gives the degrees of the irreducible factors of  $f_{364}$ , and “quad subs” gives the number of quadratic subfields of the fields defined by the degree 56 factors of  $f_{364}$ .

On our workstation—two quad-core Intel Xeon processors (2.4GHz)—our Galois group computations finished in just over 4 months (125 days). The most difficult cases (where the Galois group was either 14T9/14T29 or 14T18/14T44) took on average 20–25 hours per polynomial.

### Acknowledgements

The authors wish to thank the anonymous referee for their careful reading and useful comments, Sebastian Pauli for helpful discussions, Elon University for supporting

this project through internal grants, and the Center for Undergraduate Research in Mathematics for their support.

This research was partially funded by NSF grant #DMS-1148695.

## References

- [Amano 1971] S. Amano, “Eisenstein equations of degree  $p$  in a  $p$ -adic field”, *J. Fac. Sci. Univ. Tokyo Sect. IA Math.* **18** (1971), 1–21. MR 46 #7201 Zbl 0231.12019
- [Awtrey 2010] C. Awtrey, *Dodecic local fields*, Ph.D. thesis, Arizona State University, Tempe, AZ, 2010, available at <http://search.proquest.com/docview/305184993>. MR 2736787
- [Awtrey 2012] C. Awtrey, “Dodecic 3-adic fields”, *Int. J. Number Theory* **8:4** (2012), 933–944. MR 2926553 Zbl 1257.11101
- [Awtrey and Shill 2013] C. Awtrey and C. R. Shill, “Galois groups of degree 12 2-adic fields with automorphism group of order 6 and 12”, pp. 55–66 in *Topics from the 8th Annual UNCG Regional Mathematics and Statistics Conference* (Greensboro, NC, 2012), edited by J. Rychtář et al., Springer Proceedings in Mathematics and Statistics **64**, Springer, Heidelberg, 2013.
- [Awtrey et al.  $\geq$  2015a] C. Awtrey, B. Barkley, N. Miles, C. R. Shill, and E. Strosnider, “Computing Galois groups of degree 12 2-adic fields with trivial automorphism group”. Submitted.
- [Awtrey et al.  $\geq$  2015b] C. Awtrey, B. Barkley, N. Miles, C. R. Shill, and E. Strosnider, “Degree 12 2-adic fields with automorphism group of order 4”. To appear in *Rocky Mountain J. Math.*
- [GAP 2008] *GAP: groups, algorithms, and programming*, Version 4.4.12, The GAP Group, 2008, available at <http://www.gap-system.org>.
- [Hulpke 1999] A. Hulpke, “Techniques for the computation of Galois groups”, pp. 65–77 in *Algorithmic algebra and number theory* (Heidelberg, 1997), edited by B. H. Matzat et al., Springer, Berlin, 1999. MR 2000d:12001 Zbl 0959.12003
- [Jones and Roberts 2004] J. W. Jones and D. P. Roberts, “Nonic 3-adic fields”, pp. 293–308 in *Algorithmic number theory*, edited by D. Buell, Lecture Notes in Comput. Sci. **3076**, Springer, Berlin, 2004. MR 2006a:11156 Zbl 1125.11356
- [Jones and Roberts 2006] J. W. Jones and D. P. Roberts, “A database of local fields”, *J. Symbolic Comput.* **41:1** (2006), 80–97. MR 2006k:11230 Zbl 1140.11350
- [Jones and Roberts 2008] J. W. Jones and D. P. Roberts, “Octic 2-adic fields”, *J. Number Theory* **128:6** (2008), 1410–1429. MR 2009d:11163 Zbl 1140.11056
- [Krasner 1966] M. Krasner, “Nombre des extensions d’un degré donné d’un corps  $p$ -adique”, pp. 143–169 in *Les tendances géométriques en algèbre et théorie des nombres* (Clermont-Ferrand, 1964), edited by M. Krasner, Colloques Internationaux du Centre National de la Recherche Scientifique **143**, Éditions du CNRS, Paris, 1966. MR 37 #1349 Zbl 0143.06403
- [Lang 1994] S. Lang, *Algebraic number theory*, 2nd ed., Graduate Texts in Mathematics **110**, Springer, New York, 1994. MR 95f:11085 Zbl 0811.11001
- [Milstead et al. 2015] J. Milstead, S. Pauli, and B. Sinclair, “Constructing splitting fields of polynomials over local fields”, pp. 101–124 in *Collaborative mathematics and statistics research* (Greensboro, NC, 2013), vol. 109, edited by J. Rychtář et al., Springer Proceedings in Mathematics and Statistics **64**, Springer, Cham, 2015.
- [PARI 2012] *PARI/GP*, Version 2.5.3, The PARI Group, Bordeaux, 2012, available at <http://pari.math.u-bordeaux.fr>.
- [Pauli and Roblot 2001] S. Pauli and X.-F. Roblot, “On the computation of all extensions of a  $p$ -adic field of a given degree”, *Math. Comp.* **70:236** (2001), 1641–1659. MR 2002e:11166 Zbl 0981.11038

[Serre 1979] J.-P. Serre, *Local fields*, Graduate Texts in Mathematics **67**, Springer, New York, 1979.  
MR 82e:12016 Zbl 0423.12016

[Soicher and McKay 1985] L. Soicher and J. McKay, “Computing Galois groups over the rationals”,  
*J. Number Theory* **20**:3 (1985), 273–281. MR 87a:12002 Zbl 0579.12006

[Stauduhar 1973] R. P. Stauduhar, “The determination of Galois groups”, *Math. Comp.* **27** (1973),  
981–996. MR 48 #6054 Zbl 0282.12004

Received: 2013-08-11      Revised: 2013-08-28      Accepted: 2013-08-29

cawtre@elon.edu      *Department of Mathematics and Statistics, Elon University,  
Campus Box 2320, Elon, NC 27244, United States*

nmiles@elon.edu      *Department of Mathematics and Statistics, Elon University,  
Campus Box 3753, Elon, NC 27244, United States*

jmmilste@uncg.edu      *Department of Mathematics and Statistics,  
University of North Carolina, 116 Petty Building,  
317 College Ave, Greensboro, NC 27412, United States*

cshill@elon.edu      *Department of Mathematics and Statistics, Elon University,  
Campus Box 9017, Elon, NC 27244, United States*

estrosnider@elon.edu      *Department of Mathematics and Statistics, Elon University,  
Campus Box 5470, Elon, NC 27244, United States*

# Counting set classes with Burnside's lemma

Joshua Case, Lori Koban and Jordan LeGrand

(Communicated by Kenneth S. Berenhaut)

Mathematical tools from combinatorics and abstract algebra have been used to study a variety of musical structures. One question asked by mathematicians and musicians is: how many  $d$ -note set classes exist in a  $c$ -note chromatic universe? In the music theory literature, this question is answered with the use of Pólya's enumeration theorem. We solve the problem using simpler techniques, including only Burnside's lemma and basic results from combinatorics and abstract algebra. We use interval arrays that are associated with pitch class sets as a tool for counting.

## 1. Introduction

For the past three decades, mathematical tools from combinatorics and abstract algebra have been used to study a variety of musical structures. The elements of a  $c$ -note chromatic universe are typically labeled  $0, 1, 2, \dots, c-1$  and are considered elements of  $Z_c$ , the group of integers modulo  $c$ . In the traditional 12-note chromatic universe,  $C$  is labeled 0. Following the language of [Clough and Myerson 1985], a  $d$ -note pitch class set in a  $c$ -note chromatic universe is a subset of  $\{0, 1, \dots, c-1\}$  of size  $d$ . As explained in [Reiner 1985; Hook 2007], two pitch class sets are considered equivalent if one can be obtained from the other either by rotation or reflection. A  $d$ -note set class contains all equivalent  $d$ -note pitch class sets. One question asked by musicians and music theorists is: how many  $d$ -note set classes exist in a  $c$ -note chromatic universe? Figure 1 shows a way to visualize the case where  $c = 12$  and  $d = 7$ .

Let  $n$  be a positive integer. The Euler  $\varphi$ -function,  $\varphi(n)$ , is the number of positive integers that are less than or equal to  $n$  that are also relatively prime to  $n$ .

**Theorem 1.1** [Reiner 1985; Hook 2007]. *The number of  $d$ -note set classes in a  $c$ -note chromatic universe is*

$$\frac{1}{2c}T(c, d) + \frac{1}{2}I(c, d), \quad (1-1)$$

*MSC2010:* 00A65, 05E18.

*Keywords:* set classes, pitch class sets, Burnside's lemma, group actions.

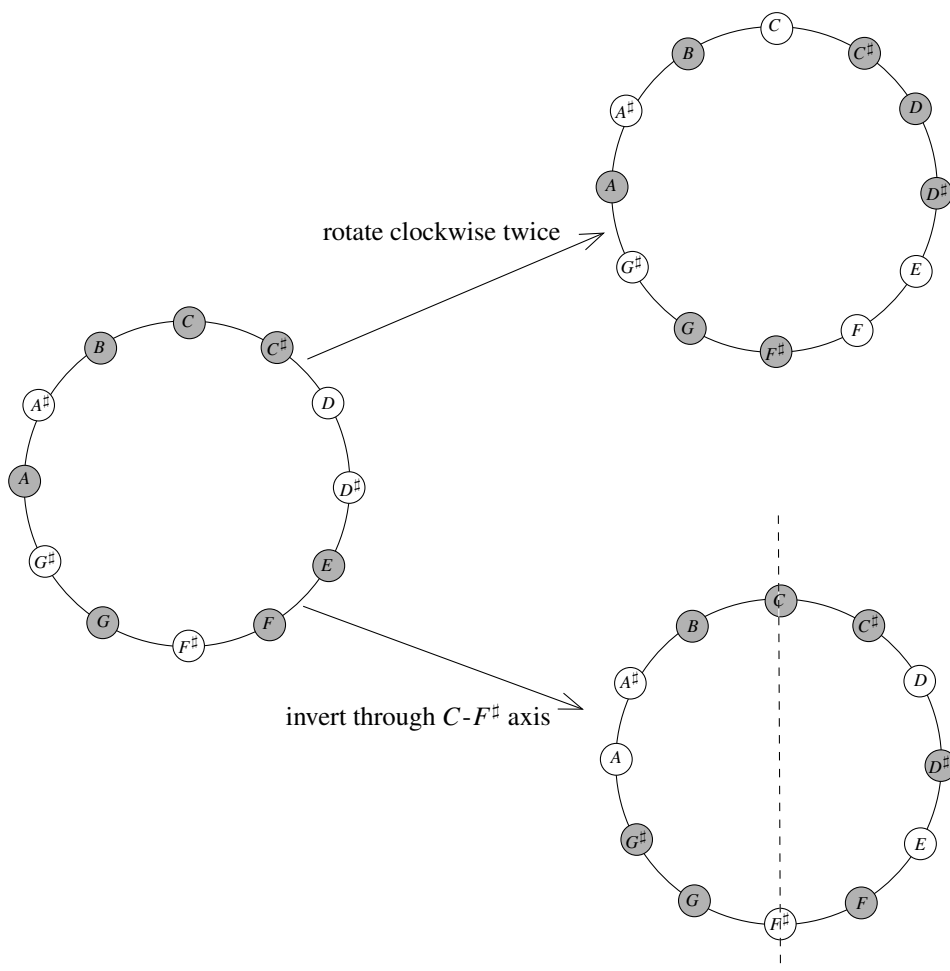
Case and LeGrand received financial support for this research from the University of Maine at Farmington's Wilson Scholar program.

where

$$T(c, d) = \sum_{j|\gcd(c,d)} \varphi(j) \binom{c/j}{d/j}$$

and

$$I(c, d) = \begin{cases} \binom{c/2-1}{\lfloor d/2 \rfloor} & \text{if } c \text{ is even and } d \text{ is odd,} \\ \binom{\lfloor c/2 \rfloor}{\lfloor d/2 \rfloor} & \text{otherwise.} \end{cases}$$



**Figure 1.** Visualizing a 7-note pitch class set in a 12-note chromatic universe. The three pitch class sets  $\{C, C^\sharp, E, F, G, A, B\}$ ,  $\{C^\sharp, D, D^\sharp, F^\sharp, G, A, B\}$ , and  $\{C, C^\sharp, D^\sharp, F, G, G^\sharp, B\}$  are equivalent and are therefore all part of the same set class.



In the music theory literature, Theorem 1.1 is proved using an advanced combinatorial theorem, namely Pólya's enumeration theorem (the final theorem stated in [Brualdi 2010]). Our contribution is that we make Theorem 1.1 more accessible by using only tools that would be seen in introductory classes in combinatorics and abstract algebra. The most advanced concept is Burnside's lemma, which appears in [Reiner 1985; Hook 2007] as a general tool for counting the number of equivalence classes generated by a group action, but is abandoned in the proof of Theorem 1.1 in favor of Pólya's result. In [Graham et al. 2008], the application of Burnside's lemma to our problem is discussed, but only specific examples, and not a general result, are reported. An additional contribution is that we use the structure of *interval arrays* (see Section 2), which were introduced in [Clough and Myerson 1985] and developed in [Fripertinger 1992], but have not been connected to this theorem.

### 2. Equivalent pitch class sets

The *dihedral group of order 2n*,  $D_{2n}$ , is the set of symmetries of a regular  $n$ -gon. There are  $n$  rotations and  $n$  reflections. Musically, rotations are known as transpositions and reflections are known as inversions.

Mathematically speaking, the number of  $d$ -note set classes in a  $c$ -note chromatic universe is the number of equivalence classes when  $D_{2c}$  acts on the set of  $d$ -note pitch class sets. In Figure 1, all 7-note pitch class sets that are equivalent to  $\{C, C^\sharp, E, F, G, A, B\}$  can be found by inverting and transposing the left-most figure in all 24 possible ways. Consult [Hook 2007] for more details about group actions in this context.

Let  $\{i_1, i_2, \dots, i_d\}$  be a  $d$ -note pitch class set. Without loss of generality, let  $i_1 < i_2 < \dots < i_d$ . The *interval array* associated with this  $d$ -note pitch class set is

$$\langle i_2 - i_1, i_3 - i_2, \dots, i_d - i_{d-1}, i_1 - i_d \rangle,$$

where all subtraction is done modulo  $d$  [Fripertinger 1992, Definition 2.5]. Note that  $\langle j_1, j_2, \dots, j_d \rangle$  is the interval array of a  $d$ -note pitch class set in a  $c$ -note chromatic universe if and only if  $j_1 + j_2 + \dots + j_d = c$  [Fripertinger 1992, Remark 2.4]. See Table 1.

Instead of counting the number of equivalence classes when  $D_{2c}$  acts on the set of  $d$ -note pitch class sets, we will count the number of equivalence classes when

7-note pitch class set	pitch class set in $Z_c$	interval array
$\{C, C^\sharp, E, F, G, A, B\}$	$\{0, 1, 4, 5, 7, 9, 11\}$	$\langle 1, 3, 1, 2, 2, 2, 1 \rangle$
$\{C^\sharp, D, D^\sharp, F^\sharp, G, A, B\}$	$\{1, 2, 3, 6, 7, 9, 11\}$	$\langle 1, 1, 3, 1, 2, 2, 2 \rangle$
$\{C, C^\sharp, D^\sharp, F, G, G^\sharp, B\}$	$\{0, 1, 3, 5, 7, 8, 11\}$	$\langle 1, 2, 2, 2, 1, 3, 1 \rangle$

**Table 1.** The interval arrays for the pitch class sets in Figure 1.

$D_{2d}$  acts on  $\{\langle j_1, j_2, \dots, j_d \rangle \mid j_1 + j_2 + \dots + j_d = c\}$ , the set of interval arrays. In Theorem 2.3 of the same work, Friperntinger proves that the number of equivalence classes is the same in both situations.

### 3. Algebraic and combinatorial tools

Below are the theorems from introductory combinatorics [Brualdi 2010] and abstract algebra [Dummit and Foote 2004] that we will apply.

**Theorem 3.1.** *Let  $n$  and  $k$  be positive integers. Then*

$$k \binom{n}{k} = n \binom{n-1}{k-1}.$$

**Theorem 3.2.** *The equation  $x_1 + x_2 + \dots + x_k = n$  has  $\binom{n-1}{k-1}$  positive-integral solutions.*

**Theorem 3.3** (hockey stick theorem). *If  $m$  and  $n$  are nonnegative integers, then*

$$\sum_{k=0}^n \binom{k}{m} = \binom{n+1}{m+1}.$$

**Theorem 3.4.** *Let  $j, k$ , and  $n$  be integers such that  $0 \leq j \leq k \leq n$ . Then*

$$\sum_{m=j}^{n-k+j} \binom{m}{j} \binom{n-m}{k-j} = \binom{n+1}{k+1}.$$

**Theorem 3.5.** *In a group, assume that element  $a$  has order  $d$ . Then*

$$\langle a^j \rangle = \langle a^{\gcd(d,j)} \rangle \quad \text{and} \quad |\langle a^j \rangle| = \frac{d}{\gcd(d,j)}.$$

**Theorem 3.6.** *If  $m$  is a positive divisor of  $d$ , then the number of elements of order  $m$  in a cyclic group of order  $d$  is  $\phi(m)$ .*

**Theorem 3.7** (Burnside's lemma). *Let  $G$  be a group acting on a set  $S$ . The number of equivalence classes is*

$$\frac{1}{|G|} \sum_{g \in G} \text{Fix}(g),$$

where  $\text{Fix}(g)$  is the number of elements of  $S$  that are fixed by  $g$ .

### 4. The main theorem proved with Burnside's lemma

**Theorem 4.1.** *The number of  $d$ -note set classes in a  $c$ -note chromatic universe is*

$$\frac{1}{2d} T_B(c, d) + \frac{1}{2} I(c, d), \tag{4-1}$$

where

$$T_B(c, d) = \sum_{m|d \text{ and } d|cm} \varphi(d/m) \binom{cm/d-1}{m-1},$$

and  $I(c, d)$  is defined as in Theorem 1.1.

*Proof.* Instead of visualizing a regular  $c$ -gon and counting the number of equivalence classes when  $D_{2c}$  acts on the set of  $d$ -note pitch class sets, as is typically done, we visualize a regular  $d$ -gon and count the number of equivalence classes when  $D_{2d}$  acts on the set of interval arrays  $\{(j_1, j_2, \dots, j_d) \mid j_1 + j_2 + \dots + j_d = c\}$ . According to Burnside's lemma, we must count the number of interval arrays that are fixed by elements of  $D_{2d}$ .

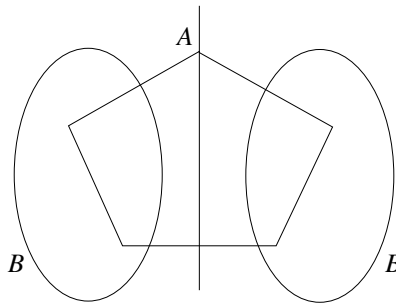
First, we consider the  $d$  inversions. Assume that  $c$  and  $d$  are both odd. We have a regular  $d$ -gon whose vertices are labeled  $j_1, j_2, \dots, j_d$ . Every possible axis of inversion passes through a single vertex. Let  $A$  be the value of that vertex, and let  $B = (c - A)/2$ . See Figure 2. Once the value of  $A$  is chosen, Theorem 3.2 says there are

$$\binom{\frac{c-A}{2} - 1}{\frac{d-1}{2} - 1}$$

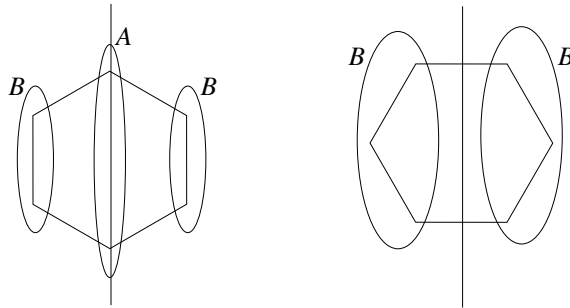
ways to assign values to the vertices that add up to  $B$ . Also note that  $A$  must be odd, and it ranges from 1 to  $c - (d - 1)$ . Thus the number of interval arrays fixed by this inversion is

$$\sum_{\substack{A=1 \\ A \text{ odd}}}^{c-(d-1)} \binom{\frac{c-A}{2} - 1}{\frac{d-1}{2} - 1},$$

which equals  $\binom{(c-1)/2}{(d-1)/2}$  by the hockey stick theorem. Since there are  $d$  inversions, the sum of the number of interval arrays fixed by an inversion is  $d \binom{\lfloor c/2 \rfloor}{\lfloor d/2 \rfloor}$ .



**Figure 2.** The inversion when  $d$  is odd.



**Figure 3.** Two inversions when  $d$  is even.

When  $c$  is even and  $d$  is odd, repeat the previous argument, except that  $A$  must be even and it ranges from 2 to  $c - (d - 1)$ . The hockey stick theorem yields

$$\binom{\frac{c-2}{2}}{\frac{d-1}{2}},$$

and the sum of the number of interval arrays fixed by an inversion is  $d \binom{c/2-1}{\lfloor d/2 \rfloor}$ .

Now assume that  $c$  and  $d$  are both even. When  $d$  is even, there are two types of inversions:  $d/2$  of each type in Figure 3. For an inversion through opposite edges, Theorem 3.2 says there are  $\binom{c/2-1}{d/2-1}$  ways to assign values to the  $d/2$  vertices that add up to  $B = c/2$ . For an inversion through a pair of vertices,  $A$  is chosen and then  $B = (c - A)/2$ . Note that  $A$  must be even and ranges from 2 to  $c - (d - 2)$ . The number of interval arrays fixed by this inversion is

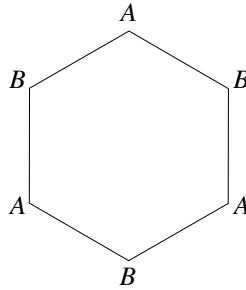
$$\begin{aligned} \sum_{\substack{A=2 \\ A \text{ even}}}^{c-(d-2)} \binom{A-1}{1} \binom{\frac{c-A}{2}-1}{\frac{d-2}{2}-1} &= \sum_{\substack{A=2 \\ A \text{ even}}}^{c-(d-2)} \binom{A}{1} \binom{\frac{c-A}{2}-1}{\frac{d-2}{2}-1} - \sum_{\substack{A=2 \\ A \text{ even}}}^{c-(d-2)} \binom{\frac{c-A}{2}-1}{\frac{d-2}{2}-1} \\ &= 2 \binom{\frac{c}{2}}{\frac{d}{2}} - \binom{\frac{c}{2}-1}{\frac{d}{2}-1}, \end{aligned}$$

where the first term simplifies by Theorem 3.4 and the second term simplifies by Theorem 3.3. The sum of the number of interval arrays fixed by the  $d$  inversions is

$$\frac{d}{2} \binom{\frac{c}{2}-1}{\frac{d}{2}-1} + d \left( 2 \binom{\frac{c}{2}}{\frac{d}{2}} - \binom{\frac{c}{2}-1}{\frac{d}{2}-1} \right) = d \binom{\frac{c}{2}}{\frac{d}{2}}.$$

The argument when  $c$  is odd and  $d$  is even is identical.

Second, we consider the  $d$  transpositions  $R^1, R^2, \dots, R^d$ , where  $R^1$  is a single transposition clockwise which generates the cyclic group of order  $d$ . Let  $m$  be a divisor of  $d$ . According to Theorem 3.5, each  $R^j$  with  $\gcd(d, j) = m$  generates the same subgroup, and this subgroup has order  $d/m$ . If an interval array can be fixed



**Figure 4.** If  $d = 6$ , rotating the hexagon  $120^\circ$  is acting on the interval arrays with  $R^2$ , an element of order 3. If an interval array is fixed, then the values  $A$  and  $B$  must each be repeated twice.

by a transposition of order  $d/m$ , it is necessary that  $(d/m) \mid c$  or, equivalently, that  $d \mid cm$ . Thus, if  $m \mid d$  and  $d \mid cm$ , the number of interval arrays fixed by an element of order  $d/m$  is the number of ordered partitions of

$$\frac{c}{d/m} = \frac{cm}{d}$$

into  $m$  parts. According to Theorem 3.2, this can be done  $\binom{cm/d-1}{m-1}$  ways. Moreover, Theorem 3.6 says that  $\varphi(d/m)$  transpositions have order  $d/m$ . Thus the sum of all  $\text{Fix}(R^j)$  is

$$\sum_{m \mid d \text{ and } d \mid cm} \varphi(d/m) \binom{cm/d-1}{m-1}.$$

See Figure 4 for an example. Applying Burnside's lemma completes the proof.  $\square$

**Theorem 4.2.** Expressions (1-1) and (4-1) are equal.

*Proof.* Since these expressions both count the number of  $d$ -note set classes in a  $c$ -note chromatic universe, they are equal. However, we provide a different proof, outside the context of music theory.

We must show that

$$\frac{1}{c} \sum_{j \mid \gcd(c,d)} \varphi(j) \binom{c/j}{d/j} = \frac{1}{d} \sum_{m \mid d \text{ and } d \mid cm} \varphi(d/m) \binom{cm/d-1}{m-1}. \tag{4-2}$$

We start with the right-hand side and reindex, letting  $j = d/m$ . Then

$$\begin{aligned} \frac{1}{d} \sum_{m \mid d \text{ and } d \mid cm} \varphi(d/m) \binom{cm/d-1}{m-1} &= \frac{1}{d} \sum_{d/j \mid d \text{ and } d \mid \frac{cd}{j}} \varphi(j) \binom{c/j-1}{d/j-1} \\ &= \frac{1}{d} \sum_{j \mid \gcd(c,d)} \varphi(j) \binom{c/j-1}{d/j-1}. \end{aligned}$$

The last equality is valid because

$$\{j : j \mid \gcd(c, d)\} = \{j : (d/j) \mid d \text{ and } d \mid (cd/j)\}.$$

The equality of (4-2) follows from termwise equality, as a result of Theorem 3.1.  $\square$

### References

- [Brualdi 2010] R. A. Brualdi, *Introductory combinatorics*, 5th ed., Pearson Prentice Hall, Upper Saddle River, NJ, 2010. MR 2012a:05001
- [Clough and Myerson 1985] J. Clough and G. Myerson, “Variety and multiplicity in diatonic systems”, *Journal of Music Theory* **29**:2 (1985), 249–270.
- [Dummit and Foote 2004] D. S. Dummit and R. M. Foote, *Abstract algebra*, 3rd ed., Wiley, Hoboken, NJ, 2004. MR 2007h:00003 Zbl 1037.00003
- [Fripertinger 1992] H. Fripertinger, “Enumeration in musical theory”, Beiträge zur elektronischen Musik 1, Hochschule für Musik und Darstellende Kunst, Graz, 1992, <http://iem.kug.ac.at/projects/workspace/projekte-bis-2008/publications/bem/bem1.html>.
- [Graham et al. 2008] J. Graham, A. Hack, and J. Wilson, “An application of Burnside’s theorem to music theory”, *The UMAP Journal* **29**:1 (2008), 45–57.
- [Hook 2007] J. Hook, “Why are there twenty-nine tetrachords? A tutorial on combinatorics and enumeration in music theory”, *Music Theory Online* **13**:4 (2007).
- [Reiner 1985] D. L. Reiner, “Enumeration in music theory”, *Amer. Math. Monthly* **92**:1 (1985), 51–54. MR 86c:05021 Zbl 0582.05005

Received: 2013-08-14    Revised: 2013-10-24    Accepted: 2013-12-23

joshua.case@maine.edu	<i>Mathematics Department, University of Maine at Farmington, 228 South Street, Farmington, ME 04938, United States</i>
lori.koban@maine.edu	<i>Mathematics Department, University of Maine at Farmington, 228 South Street, Farmington, ME 04938, United States</i>
jordan.legrand@maine.edu	<i>Mathematics Department, University of Maine at Farmington, 228 South Street, Farmington, ME 04938, United States</i>

# Border rank of ternary trilinear forms and the $j$ -invariant

Derek Allums and Joseph M. Landsberg

(Communicated by David Royal Larson)

We first describe how one associates a cubic curve to a given ternary trilinear form  $\phi \in \mathbb{C}^3 \otimes \mathbb{C}^3 \otimes \mathbb{C}^3$ . We explore relations between the rank and border rank of the tensor  $\phi$  and the geometry of the corresponding cubic curve. When the curve is smooth, we show there is no relation. When the curve is singular, normal forms are available, and we review the explicit correspondence between the normal forms, rank and border rank.

## 1. Introduction

Given a multilinear map, i.e., a tensor<sup>1</sup>, how hard is it to evaluate? Two ways mathematicians have chosen to quantify “hard” are the notions of rank and border rank. We say a tensor  $\phi \in V_1 \otimes \cdots \otimes V_n$  is of *rank 1* if it is of the form  $v_1 \otimes \cdots \otimes v_n$ , where each  $v_i \in V_i$ .

**Definition 1.1.** Let  $\phi \in V_1 \otimes \cdots \otimes V_n$ . The rank of  $\phi$ , denoted  $\mathbf{R}(\phi)$  is the smallest natural number  $r$  such that  $\phi = \sum_{j=1}^r \phi_j$ , where each  $\phi_j \in V_1 \otimes \cdots \otimes V_n$  is of rank 1.

To better understand this concept, consider the reduction to linear algebra, in which  $\phi \in V_1 \otimes V_2$  may be considered as a linear map  $V_1^* \rightarrow V_2$ . Recall that every linear map on finite dimensional vector spaces can be written as a matrix, after choosing bases, and that the rank of a matrix  $M$  is the number of rank 1 matrices  $M_i$  needed to write  $M = \sum_i M_i$ . In this special case, the above definition is natural.<sup>2</sup>

But rank doesn’t give us the whole picture when  $n > 2$ . To illustrate this, consider the following classical example.

---

*MSC2010:* 15A72, 68Q17.

*Keywords:* algebraic geometry, border rank of tensors,  $j$ -invariant of cubic, ternary trilinear forms.

<sup>1</sup>Throughout the paper, we will assume the reader is familiar with the tensor product of vector spaces. For a quick review, see the Appendix.

<sup>2</sup>However, it is worth mentioning that *rank* as it is defined here is one of several generalizations of the rank of a linear map (e.g., multilinear rank).

The tensor

$$\phi = a_1 \otimes b_1 \otimes c_1 + a_1 \otimes b_1 \otimes c_2 + a_1 \otimes b_2 \otimes c_1 + a_2 \otimes b_1 \otimes c_1$$

is of rank at most 3 since

$$\phi = a_1 \otimes b_1 \otimes (c_1 + c_2) + a_1 \otimes b_2 \otimes c_1 + a_2 \otimes b_1 \otimes c_1,$$

and it is not of rank 2 by explicit computation. However, notice that  $\phi$  is the limit as  $\epsilon \rightarrow 0$  of the following sequence of rank 2 tensors:

$$\phi(\epsilon) = \frac{1}{\epsilon}((\epsilon - 1)a_1 \otimes b_1 \otimes c_1 + (a_1 + \epsilon a_2) \otimes (b_1 + \epsilon b_2) \otimes (c_1 + \epsilon c_2)).$$

So the rank of the tensor is 3, but we can approximate it as closely as we like with rank 2 tensors. We say  $\phi$  has *border rank 2*, and we have the following definition.

**Definition 1.2.** A tensor  $\phi \in V_1 \otimes \cdots \otimes V_n$  is said to be of border rank  $r$ , denoted  $\mathbf{R}(\phi) = r$ , if it is the limit of tensors of rank  $r$  but not of tensors of rank  $s$  for any  $s < r$ .

One way to approach the difficult general problem of understanding the border rank of tensors is to reduce multilinear algebra to linear algebra. Below is one such reduction, in which we consider  $\phi \in A \otimes B \otimes C = \mathbb{C}^3 \otimes \mathbb{C}^3 \otimes \mathbb{C}^3$  as a linear map  $A^* \rightarrow B \otimes C$  and then represent the image in  $B \otimes C$  as a matrix. We then take the determinant of this representation to find an associated cubic curve to  $\phi$ .

Choose bases  $\{a_i\}, \{b_j\}, \{c_k\}$  for  $A, B, C$ , respectively, with  $\{a_i^*\}, \{b_j^*\}, \{c_k^*\}$  the dual bases. Now let

$$\phi = \sum_{i,j,k} \phi_{ijk} a_i \otimes b_j \otimes c_k \in A \otimes B \otimes C,$$

where  $\phi_{ijk} \in \mathbb{C}$  are constants and let

$$a^* = xa_i^* + ya_j^* + za_k^*, \quad x, y, z \in \mathbb{C},$$

be an arbitrary element of  $A^* = (\mathbb{C}^3)^*$ . Then, the matrix representation of  $\phi$  parametrized by  $a^*$ , denoted  $[\phi \lrcorner a^*]$ , has  $(j, k)$ -th entry

$$[\phi \lrcorner a^*]_{j,k} = \phi_{1jk}x + \phi_{2jk}y + \phi_{3jk}z.$$

In the same way, we can find matrix representations  $[\phi \lrcorner b^*]$  and  $[\phi \lrcorner c^*]$  parametrized by  $b^* \in B^*$  and  $c^* \in C^*$ . For the tensors we study in this paper, all of these representations turn out to be equal, so we work with  $[\phi \lrcorner a^*]$  without loss of generality.

Let's look at an example. If

$$\phi = a_1 \otimes b_1 \otimes c_1 + a_2 \otimes b_2 \otimes c_2 + a_3 \otimes b_1 \otimes c_2 + a_3 \otimes b_3 \otimes c_3,$$

then,

$$\phi_{111} = \phi_{222} = \phi_{312} = \phi_{333} = 1,$$



and  $\phi_{ijk} = 0$  otherwise. Thus,

$$[\phi \lrcorner a^*] = \begin{pmatrix} x & z & 0 \\ 0 & y & 0 \\ 0 & 0 & z \end{pmatrix}.$$

Now take the determinant to find the determinantal cubic associated to  $\phi$ ,

$$xyz = 0.$$

It has been known since as early as 1938 (see e.g., [Thrall and Chanler 1938]) that any cubic curve in three variables is projectively equivalent to one of the following:

- (1) triple line  $x^3 = 0$
- (2) double line and a line  $x^2y = 0$
- (3) 3 lines intersecting at a point  $xy(x - y) = 0$
- (4) 3 lines in general position  $xyz = 0$
- (5) a conic and a tangent line  $z(x^2 + yz) = 0$
- (6) a conic and a transverse line  $x(x^2 + yz) = 0$
- (7) cuspidal cubic  $x^3 - y^2z = 0$
- (8) node  $x^3 + y^3 - xyz = 0$
- (9) a smooth cubic: the general case
- (10) a cubic identically zero

The tensors to which these other singular cases correspond are dealt with in [Thrall and Chanler 1938] and later in more modern language in [Ng 1995]. In particular, normal forms are given, and in [Allums 2011], the border rank of each of these singular tensors is calculated.

Since the singular cases have been dealt with, the next question is: how is border rank related to the intrinsic geometry of the determinantal cubic in the general case? That is, how does the border rank vary in the open set of *smooth* cubics? To answer this, we need to introduce the classical invariants  $S$ ,  $T$  and  $J$ , which are rational functions in the coefficients of a cubic.

Under the action of  $SL(\mathbb{C}, 3)$  on the cubic, there is a unique (up to scale) degree 4 invariant  $S$  and a unique (up to scale) degree 6 invariant  $T$  [Sturmfels 1993]. These generate the ring of invariants of a cubic of which

$$J := \frac{S^3}{T^2 - 64S^3},$$

the  $j$ -invariant, is a member. The invariants  $S$  and  $T$  are extrinsic invariants of the curve, while  $J$  is an intrinsic invariant<sup>3</sup>. Here this means  $S$  and  $T$  classify the curve up to change of coordinates while  $J$  classifies smooth cubics up to isomorphism as abelian varieties, i.e., as groups and as algebraic varieties. One goal of this paper is to find out what relationship, if any, exists between the border rank of  $\phi \in \mathbb{C}^3 \otimes \mathbb{C}^3 \otimes \mathbb{C}^3$  and the geometry of its determinantal cubic curve. Equivalently, we want to describe the relationship between border rank and  $S$ ,  $T$  and thus  $J$ .

The maximum possible border rank of  $\phi \in \mathbb{C}^3 \otimes \mathbb{C}^3 \otimes \mathbb{C}^3$  is 5 [Landsberg 2012], and since a tensor of border rank 5 depends on twelve parameters, we start with a smaller case and consider tensors of border rank 4, which we show depend on only three parameters in Proposition 3.1. We take such a tensor and calculate the invariants  $S$  and  $T$  of its determinantal cubic, summarizing our analysis in Proposition 3.2. In particular, we conclude that there is no meaningful relationship between the border rank of  $\phi$  and  $S$  or  $T$ , and thus no meaningful relationship between border rank and  $J$ , if the cubic is smooth.

## 2. Background

Some background material is given in the appendix. We present the rest here, with most of it coming from [Landsberg 2012].

There exists a geometric interpretation of border rank as follows. Let  $V$  be a finite dimensional complex vector space and let  $X \subset \mathbb{P}V$  be a variety. For any point  $q$  not on  $X$ , we define the *join of  $q$  and  $X$*  to be the set of all secant lines containing  $q$  and some point of  $X$ , denoted  $J(q, X)$ . If  $q = x \in X$ , we do the same thing, but we also allow tangent lines at  $x$  since a tangent line is a limit of secant lines. The *secant variety* of  $X$  is

$$\sigma(X) := \overline{\bigcup_{x \in X} J(x, X)},$$

where the bar denotes Zariski closure. The notation  $J(X, X) = \sigma(X)$  is also used. We can also define the join of two distinct varieties  $Y, Z \subset \mathbb{P}V$  by

$$J(Y, Z) = \overline{\bigcup_{q \in Y} J(q, Z)},$$

where  $J(q, Z)$  is the set of all secant lines containing  $q \in Y$  and some point of  $Z$ .

**Definition 2.1** [Landsberg 2012]. The join of  $k$  varieties  $X_1, \dots, X_k \subset \mathbb{P}V$  is the closure of the union of the corresponding secant  $(k-1)$ -planes, or by induction,

---

<sup>3</sup>Consider the difference between “extrinsic” and “intrinsic” in surface theory: mean curvature is extrinsic (invariant under Euclidean motion) but Gauss curvature is intrinsic (invariant under isometry).

$J(X_1, \dots, X_k) = J(X_1, J(X_2, \dots, X_k))$ . Define the  $k$ -th secant variety of  $X$  to be  $\sigma_k(X) = J(X, \dots, X)$ , the join of  $k$  copies of  $X$ .

We move on to another crucial concept: the Segre variety.

**Definition 2.2.** The  $n$ -factor Segre variety is the image of the map

$$\begin{aligned} \text{Seg} : \mathbb{P}V_1 \times \dots \times \mathbb{P}V_n &\rightarrow \mathbb{P}(V_1 \otimes \dots \otimes V_n), \\ ([v_1], \dots, [v_n]) &\mapsto [v_1 \otimes \dots \otimes v_n]. \end{aligned}$$

Note that for fixed  $n \in \mathbb{N}$ , the image of the Segre map is the projectivization of the rank 1  $n$ -tensors.

A tensor  $\phi \in V_1 \otimes \dots \otimes V_n$  may be interpreted as a linear map

$$V_1^* \rightarrow V_2 \otimes \dots \otimes V_n, \dots, V_n^* \rightarrow V_1 \otimes \dots \otimes V_{n-1}.$$

Recall a matrix is rank 1 if and only if all its  $2 \times 2$  minors are 0. The set of rank 1 tensors in  $V_1 \otimes \dots \otimes V_n$  is exactly the set of tensors such that each of the previous linear maps has rank 1 [Landsberg 2012]. The collection of these  $2 \times 2$  minors are homogeneous polynomials called flattenings. Thus, using Definition 5.3, the set of tensors of rank 1 is an algebraic variety.

Tensors of border rank  $r$  are described as limits of tensors of rank  $r$ , so the set of tensors of border rank at most  $r$  is the closure of the set of tensors of rank  $r$ , where a tensor of rank  $r$  is contained in the linear span of  $r$  points of the set of tensors of rank 1. Since in this case the Zariski and Euclidean closures coincide (see [Mumford 1976, Theorem 2.33]), the (projectivization of the) set of tensors of border rank at most  $r$  is thus exactly  $\sigma_r(\text{Seg}(\mathbb{P}V_1 \times \dots \times \mathbb{P}V_n))$ , and so we now have an entirely geometric interpretation of border rank with which to work. In particular, we can now restate some of the introduction in more modern language.

For  $A \otimes B \otimes C = \mathbb{C}^3 \otimes \mathbb{C}^3 \otimes \mathbb{C}^3$ , the representation of  $\phi$  as a matrix defines a vector space of matrices in  $\phi(A^*) \subset B \otimes C$  of dimension 3 parametrized by  $a^* \in A^*$ . When we move into projective space, it becomes a copy of  $\mathbb{P}^2 \subset \mathbb{P}(B \otimes C)$ . By requiring that its determinant vanish, we are demanding that the matrix be of rank at most 2. That is, we want the matrix to be contained in  $\sigma_2(\text{Seg}(\mathbb{P}B \times \mathbb{P}C))$ . Our goal is then to see how border rank varies in the intersection

$$\{\mathbb{P}(\phi(A^*)) \mid \phi \in A \otimes B \otimes C\} \cap \sigma_2(\text{Seg}(\mathbb{P}B \times \mathbb{P}C)).$$

### 3. Primary results

First, we show that a general point in  $\sigma_4 := \sigma_4(\text{Seg}(\mathbb{P}A \times \mathbb{P}B \times \mathbb{P}C))$ , i.e., a tensor of border rank 4, depends on only three parameters.

**Proposition 3.1.** *A general point in  $\sigma_4$ , up to the action of  $\text{GL}(\mathbb{C}, 3)$ , depends on exactly three parameters.*

*Proof.* Let  $\zeta_i, \alpha_i, \beta_i, \gamma_i \in \mathbb{C}$  be constants and choose bases  $\{a_i\}, \{b_i\}, \{c_i\}$  for  $A, B, C$ . We first show that  $a_1 \otimes b_1 \otimes c_1 + a_2 \otimes b_2 \otimes c_2 + a_3 \otimes b_3 \otimes c_3$  is a general point in  $\sigma_3$  by beginning with an arbitrary general point in  $\sigma_3$ . To do this, define

$$\begin{aligned} u_i &= \alpha_{i1} a_1 + \alpha_{i2} a_2 + \alpha_{i3} a_3, \\ v_j &= \beta_{j1} b_1 + \beta_{j2} b_2 + \beta_{j3} b_3, \\ w_k &= \gamma_{k1} c_1 + \gamma_{k2} c_2 + \gamma_{k3} c_3, \end{aligned}$$

where  $\alpha_{ip}, \beta_{jp}, \gamma_{kp}$  are constants such that each set  $\{u_i\}, \{v_j\}, \{w_k\}$  is linearly independent, which can be done in any open set; so this is a sufficiently arbitrary choice of elements. Let

$$u_1 \otimes v_1 \otimes w_1 + u_2 \otimes v_2 \otimes w_2 + u_3 \otimes v_3 \otimes w_3$$

be a general point in  $\sigma_3$ . Since our group of normalizations,  $\text{GL}(\mathbb{C}, 3)$ , is 9-dimensional, we can send each  $u_i \mapsto a_i, v_j \mapsto b_j$  and  $w_k \mapsto c_k$ , totaling nine transformations. We then have

$$a_1 \otimes b_1 \otimes c_1 + a_2 \otimes b_2 \otimes c_2 + a_3 \otimes b_3 \otimes c_3, \tag{11}$$

as desired. A general point in  $\sigma_4$  is obtained by taking an arbitrary point in  $\text{Seg}(\mathbb{P}A \times \mathbb{P}B \times \mathbb{P}C)$  and adding it to (11) to obtain a point on an honest secant line:

$$\begin{aligned} &a_1 \otimes b_1 \otimes c_1 + a_2 \otimes b_2 \otimes c_2 + a_3 \otimes b_3 \otimes c_3 \\ &+ (\alpha_1 a_1 + \alpha_2 a_2 + \alpha_3 a_3) \otimes (\beta_1 b_1 + \beta_2 b_2 + \beta_3 b_3) \otimes (\gamma_1 c_1 + \gamma_2 c_2 + \gamma_3 c_3). \end{aligned}$$

Since  $\text{GL}(\mathbb{C}, 3)$  is 9-dimensional, we may make six dimensions worth of changes by sending  $\alpha_i a_i \mapsto a_i$  and  $\beta_j b_j \mapsto b_j$ , with three dimensions worth of changes left over. However, these transformations add additional constants to the first three summands; we end up with

$$\sum_{i=1}^3 \frac{1}{\alpha_i \beta_i} a_i \otimes b_i \otimes c_i + (a_1 + a_2 + a_3) \otimes (b_1 + b_2 + b_3) \otimes (\gamma_1 c_1 + \gamma_2 c_2 + \gamma_3 c_3).$$

Using our last three dimensions to send

$$\frac{1}{\alpha_i \beta_i} c_i \mapsto c_i$$

gives

$$\sum_{i=1}^3 a_i \otimes b_i \otimes c_i + (a_1 + a_2 + a_3) \otimes (b_1 + b_2 + b_3) \otimes (\alpha_1 \beta_1 \gamma_1 c_1 + \alpha_2 \beta_2 \gamma_2 c_2 + \alpha_3 \beta_3 \gamma_3 c_3).$$

Finally, for the sake of notation, relabel

$$\lambda_i = \alpha_i \beta_i \gamma_i.$$

Thus, a general point in  $\sigma_4$ ,

$$a_1 \otimes b_1 \otimes c_1 + a_2 \otimes b_2 \otimes c_2 + a_3 \otimes b_3 \otimes c_3 + (a_1 + a_2 + a_3) \otimes (b_1 + b_2 + b_3) \otimes (\lambda_1 c_1 + \lambda_2 c_2 + \lambda_3 c_3),$$

depends on only the three parameters  $\lambda_1, \lambda_2, \lambda_3$ . □

Note that the action of  $GL(\mathbb{C}, 3)$  on  $\sigma_4$  does not change  $S$  or  $T$  as these are invariant under changes of coordinates. Now represent this tensor as a matrix, as described in the introduction:

$$\begin{pmatrix} x + \lambda_1(x + y + z) & \lambda_2(x + y + z) & \lambda_3(x + y + z) \\ \lambda_1(x + y + z) & y + \lambda_2(x + y + z) & \lambda_3(x + y + z) \\ \lambda_1(x + y + z) & \lambda_2(x + y + z) & z + \lambda_3(x + y + z) \end{pmatrix}.$$

Take the determinant to find the determinantal cubic curve, which is

$$(1 + \gamma_1 + \gamma_2 + \gamma_3)xyz + \gamma_1y^2z + \gamma_1yz^2 + \gamma_2x^2z + \gamma_2xz^2 + \gamma_3x^2y + \gamma_3xy^2. \tag{12}$$

From here, one uses the formulae for  $S$  and  $T$  found in [Sturmfels 1993].

**Proposition 3.2.** *The border rank of  $\phi \in \mathbb{C}^3 \otimes \mathbb{C}^3 \otimes \mathbb{C}^3$  is not related to the projective geometry of its determinantal cubic curve, if it is smooth.*

*Proof.* The polynomials  $S$  and  $T$  are in the ten coefficients of a cubic in general, but as shown in Proposition 3.1, the coefficients of our curve depends only on three parameters  $\gamma_1, \gamma_2, \gamma_3$ , so here  $S$  and  $T$  are in three variables. Now fix  $\gamma_1 = \gamma_2 = 1$ . Then  $S$  and  $T$  become nonconstant polynomials in the single complex variable  $\gamma_3$ :

$$S = \frac{1}{16}\gamma_3^4 - \frac{5}{12}\gamma_3^3 + \frac{7}{8}\gamma_3^2 + \frac{43}{108}\gamma_3 + \frac{169}{1296},$$

$$T = -\frac{1}{8}\gamma_3^6 + \frac{5}{4}\gamma_3^5 - \frac{113}{24}\gamma_3^4 + \frac{283}{54}\gamma_3^3 + \frac{691}{216}\gamma_3^2 - \frac{559}{324}\gamma_3 - \frac{2197}{5832}.$$

By Picard’s theorem,  $S$  and  $T$  each either attain every value in  $\mathbb{C}$  or attain all but one value in  $\mathbb{C}$ . However, if there was some  $w \in \mathbb{C}$  not hit by  $S$  or  $T$ , then  $S = w$  would have no solution. But since  $\mathbb{C}$  is algebraically closed,  $S - w = 0$  does have a root. Thus,  $S$  and  $T$  are onto, so we may obtain any value for them by suitable choices of  $\gamma_1, \gamma_2, \gamma_3$ . □

#### 4. On the 24 singular cases

Define

$$\Delta := T^2 - 64S^3$$

to be the *discriminant* of a cubic curve. Since a cubic is singular if and only if  $\Delta = 0$ , one expects each of the determinantal cubics associated to the normal forms in [Ng 1995] to have  $\Delta = 0$ . The determinantal cubics are:

$$xyz = 0 \quad \{1, 2, 3, 5, 6, 8\}$$

$$xyz - x^3 = 0 \quad \{4, 9, 10\}$$

$$(\lambda - 1)xyz = 0 \quad \{7\}$$

$$y^2z + yz^2 = 0 \quad \{11\}$$

$$x^2y + xy^2 = 0 \quad \{12\}$$

$$x^2y - xz^2 = 0 \quad \{13, 14\}$$

$$(\lambda - 1)(\lambda z^3 + xyz) = 0 \quad \{15\}$$

$$xyz - \lambda z^3 + y^3 = 0 \quad \{16\}$$

$$xyz + \lambda x^3 = 0 \quad \{17, 18\}$$

$$z^2y - zy^2 - xy^2 = 0 \quad \{19\}$$

$$xz^2 + y^3 + \mu zy^2 = 0 \quad \{20\}$$

$$-\mu x^2y - xy^2 + x^2z = 0 \quad \{21, 22\}$$

$$\begin{aligned} &(\lambda_3\lambda_5)z^3 + (\lambda_1\lambda_5 + \lambda_4\lambda_6)xz^2 \\ &+ (\lambda_2\lambda_6)y^2z + (\lambda_2\lambda_5 + \lambda_3\lambda_6)yz^2 \\ &\quad - (\lambda_4\lambda_6 + \lambda_1\lambda_5)xy^2 + (\lambda_1\lambda_6)xyz = 0 \end{aligned} \quad \{23\}$$

$$-\mu z^3 - 2\mu^3y^2z + 3\mu^2yz^2 + 3\mu xy^2 = 0 \quad \{24\}$$

The set of numbers to the right are the normal forms to which the curve corresponds and

$$\lambda_1 = (\lambda - 1), \quad \lambda_2 = (\lambda - 1)^2(\lambda^2 + \lambda + 1), \quad \lambda_3 = (\lambda^2 - 1)(\lambda^2 + \lambda + 1),$$

$$\lambda_4 = (\lambda + 1), \quad \lambda_5 = (\lambda^2 + 1), \quad \lambda_6 = (\lambda^2 - 1),$$

where  $\lambda \neq 0, 1$  for  $\{7, 15\}$ ;  $\lambda \neq 0$  for  $\{16, 17, 18\}$ ;  $\lambda \neq 0, \omega$  for  $\{23\}$  (where  $\omega^3 = 1$ );  $\mu = 0, 1$  for  $\{20, 21, 22\}$ ; and  $\mu \neq 0$  for  $\{24\}$ . Using the formulae in [Sturmfels 1993], we find  $\Delta = 0$  for each of these cubics.

Notice that some of these cubics are projectively equivalent. Some of these equivalences are immediate<sup>4</sup>, such as

$$\{1, 2, 3, 5, 6, 8\}, \{7\} \sim (4),$$

$$\{4, 9, 10\}, \{15\}, \{17, 18\} \sim (6),$$

$$\{11\}, \{12\} \sim (3),$$

$$\{16\} \sim (8),$$

<sup>4</sup>Explanation of notation by example: The cubics  $\{1, 2, 3, 5, 6, 8\}$  in [Ng 1995] correspond to  $xyz = 0$  above, and this corresponds to three lines in general position, which is case (4) in [Thrall and Chanler 1938]. Additionally,  $\{7\}$  corresponds to  $(\lambda - 1)xyz = 0$ , which is projectively equivalent to  $xyz = 0$  and so (4) as well. Thus we write  $\{1, 2, 3, 5, 6, 8\}, \{7\} \sim (4)$ .

where the numbers to the right come from the classification in the introduction. To find the others, we find the singular points and expand in a Taylor series about that point. We then look at the second order term: if it is of rank 1, then the singularity is a cusp, and if it is of rank 2, the singularity is a node. As an example, let's examine  $f(x, y, z) = x^2y - xz^2$ , which is the cubic corresponding to {13, 14}. The curve is singular at a point  $p$  if and only if the differential,  $D$ , vanishes at  $p$ . In this case,

$$D = (2xy - z^2, x^2, -2xz).$$

Since  $D(p) = 0$  if and only if  $p = [x : y : z] = [0 : 1 : 0]$ , this is our singular point. Expand in a Taylor series about this point:

$$f(x, y, z) = f(p) + xf_x(p) + yf_y(p) + zf_z(p) + \frac{1}{2}x^2 f_{xx}(p) + \dots .$$

The only nonzero term of second order is  $\frac{1}{2}x^2 f_{xx}(p) = x^2$ , which is of rank 1. Thus, our curve has a cusp and corresponds to case (7).

The classification of the remaining cases is a simple exercise in calculus, and we end up with

$$\begin{aligned} \{13, 14\}, \{19\}, \{20\}, \{21, 22\}, \{24\} &\sim (7), \\ \{23\} &\sim (8). \end{aligned}$$

### 5. Appendix

We begin with the definition of the tensor product of vector spaces. Although the tensor product is typically defined by its universal property, those familiar with it will have no trouble relating the following definition, which is sufficient for our purposes, to the standard one. In all cases,  $\otimes = \otimes_{\mathbb{C}}$  and recall that for a vector space  $V$ , we denote by  $V^*$  the dual space to  $V$ , which is the space of all linear maps  $V \rightarrow \mathbb{C}$ .

**Definition 5.1.** Let  $V_1, \dots, V_n, W$  be finite-dimensional vector spaces. A map  $f : V_1 \times \dots \times V_n \rightarrow W$  is said to be  $n$ -linear if it is linear in each factor. The tensor product of these spaces is

$$V_1 \otimes \dots \otimes V_n \otimes W = \{f : V_1^* \times \dots \times V_n^* \rightarrow W \mid f \text{ is } n\text{-linear}\}.$$

Note that when  $W = \mathbb{C}$ , we have that

$$V_1 \otimes \dots \otimes V_n \otimes W = V_1 \otimes \dots \otimes V_n \otimes \mathbb{C} \simeq V_1 \otimes \dots \otimes V_n.$$

This is a standard result, whose statement in full generality can be seen in, e.g., Theorem 5.7 in [Hungerford 1980]. It is a straightforward exercise to show that  $V \otimes W$  is the space of linear maps  $V^* \rightarrow W$ , the space of linear maps  $W^* \rightarrow V$ , the space of bilinear maps  $V^* \times W^* \rightarrow \mathbb{C}$ , etc. Inductively, we have many different equivalent ways to realize  $V_1 \otimes \dots \otimes V_n \otimes W$ . The tensor product of vector spaces is again a vector space, whose elements are called tensors.

Next, since our work is done in complex projective space, we need a definition;  $n$ -dimensional complex projective space is the space of all one-dimensional subspaces (lines) in  $\mathbb{C}^{n+1}$  [Harris 1995]:

**Definition 5.2.** Define  $n$ -dimensional complex projective space to be

$$\mathbb{P}^n = \mathbb{P}\mathbb{C}^n := (\mathbb{C}^{n+1} \setminus \{0\}) / \sim,$$

where  $\sim$  is the equivalence relation given by  $\mathbb{C}^n \ni (v_1, \dots, v_n) \sim (\lambda v_1, \dots, \lambda v_n)$  for some nonzero scalar  $\lambda$ .

For a complex vector space  $V$  of finite dimension, denote the set of equivalence classes of some  $v \in V$  by  $[v] \in \mathbb{P}V$ . Let

$$\begin{aligned} \pi : V \setminus \{0\} &\rightarrow \mathbb{P}V, \\ v &\mapsto [v] \end{aligned}$$

denote the projection. For a subset  $Z \subset \mathbb{P}V$ , let  $\hat{Z} := \pi^{-1}(Z)$  denote the cone over  $Z$ . Call the image of such a cone in projective space its projectivization. We need a final crucial definition from [Harris 1995]:

**Definition 5.3.** A projective variety is the projectivization of the set of common zeros of some collection of homogeneous polynomials on  $V$ .

Should the reader want to read more relevant background material, see the sections on the tensor product in [Landsberg 2012; Hungerford 1980; Dummit and Foote 2004] and the sections on basic algebraic geometry in [Landsberg 2012; Harris 1995].

## References

- [Allums 2011] D. J. Allums, *Toward a classification of the ranks and border ranks of all (3,3,3) trilinear forms*, junior thesis, Texas A&M University, 2011, available at <http://repository.tamu.edu/bitstream/handle/1969.1/ETD-TAMU-2011-05-9621/ALLUMS-THESIS.pdf?sequence=2>.
- [Dummit and Foote 2004] D. S. Dummit and R. M. Foote, *Abstract algebra*, 3rd ed., John Wiley & Sons, Hoboken, NJ, 2004. MR 2007h:00003 Zbl 1037.00003
- [Harris 1995] J. Harris, *Algebraic geometry: a first course*, Graduate Texts in Mathematics **133**, Springer, New York, 1995. MR 97e:14001
- [Hungerford 1980] T. W. Hungerford, *Algebra*, Graduate Texts in Mathematics **73**, Springer, New York, 1980. MR 82a:00006 Zbl 0442.00002
- [Landsberg 2012] J. M. Landsberg, *Tensors: geometry and applications*, Graduate Studies in Mathematics **128**, American Mathematical Society, Providence, RI, 2012. MR 2865915 Zbl 1238.15013
- [Mumford 1976] D. Mumford, *Algebraic geometry, I: Complex projective varieties*, Grundlehren der Math. Wissenschaften **221**, Springer, Berlin, 1976. MR 56 #11992 Zbl 0356.14002
- [Ng 1995] K. O. Ng, “The classification of (3, 3, 3) trilinear forms”, *J. Reine Angew. Math.* **468** (1995), 49–75. MR 97a:14051 Zbl 0858.11023



[Sturmfels 1993] B. Sturmfels, *Algorithms in invariant theory*, Springer, Vienna, 1993 MR 94m:13004  
Zbl 0802.13002

[Thrall and Chanler 1938] R. M. Thrall and J. H. Chanler, "Ternary trilinear forms in the field of  
complex numbers", *Duke Math. J.* 4:4 (1938), 678–690. MR 1546088 Zbl 0020.06105

Received: 2013-09-18      Accepted: 2014-01-24

derek.allums@rice.edu

*Department of Mathematics, Rice University,  
Houston, TX 77005, United States*

jml@math.tamu.edu

*Texas A&M University, College Station, TX 77843,  
United States*



# On the least prime congruent to 1 modulo $n$

Jackson S. Morrow

(Communicated by Kenneth S. Berenhaut)

For any integer  $n > 1$ , there are infinitely many primes congruent to 1 (mod  $n$ ). In this note, the elementary argument of Thangadurai and Vatwani is modified to improve their upper estimate of the least such prime when  $n$  itself is a prime greater than or equal to 5.

## Preliminaries

For any integer  $n \geq 1$ , the  $n$ -th cyclotomic polynomial is

$$\Phi_n(x) = \prod_{\substack{1 \leq m \leq n \\ \gcd(m,n)=1}} (x - e^{2\pi im/n}).$$

This is a monic polynomial of degree  $\varphi(n)$ , where  $\varphi$  denotes Euler's phi function, and the roots of this polynomial are the primitive complex  $n$ -th roots of unity. It is well-known that  $\Phi_n(x)$  is irreducible over  $\mathbb{Q}$ , with integer coefficients, and  $x^n - 1 = \prod_{d|n} \Phi_d(x)$ . From the last equation, we have

$$\Phi_n(x) = \frac{x^n - 1}{\prod_{\substack{d|n \\ d < n}} \Phi_d(x)}. \quad (1)$$

It is a consequence of a well-known result of Dirichlet [1889] that for each integer  $n > 0$ , there are infinitely many primes of the form  $kn + 1$ , where  $k$  is a positive integer. The problem of determining, or estimating, the smallest prime  $p^*(n) \equiv 1 \pmod{n}$  has attracted interest. In [Heath-Brown 1992; Linnik 1944a; 1944b; Xylouris 2009], estimates of the form  $p^*(n) \leq c_1 n^{c_2}$ , with  $c_1, c_2$  constants independent of  $n$ , are proven using highly nonelementary methods of analytic number theory. Recently, elementary proofs of weaker bounds on  $p^*(n)$  have been given. In [Sabia and Tesauri 2009], it is shown that  $p^*(n) \leq (3^n - 1)/2$ ; in [Thangadurai and Vatwani 2011], this is improved to  $p^*(n) \leq 2^{\varphi(n)+1} - 1$ . Here

*MSC2010:* 11B25, 11N13.

*Keywords:* primes in progressions, arithmetic progressions.

This work was supported by NSF grant no. 1262930, and was completed during the 2013 Research Experience for Undergraduates Program in Algebra and Discrete Mathematics at Auburn University.

we adapt the methods of [Thangadurai and Vatwani 2011] (which were adapted from [Sabia and Tesauri 2009]) to prove the following theorem.

**Theorem.** *Let  $n \geq 5$  be a prime. The smallest prime  $p^*(n) \equiv 1 \pmod{n}$  satisfies the bound*

$$p^*(n) \leq (2^n + 1)/3.$$

### Main result

From (1), we see that if  $n$  is a prime, then

$$\Phi_n(X) = \frac{X^n - 1}{X - 1} = X^{n-1} + \dots + 1, \quad (2)$$

and if  $n$  is an odd prime,

$$\begin{aligned} \Phi_{2n}(X) &= \frac{X^{2n} - 1}{\Phi_1(X)\Phi_2(X)\Phi_n(X)} = \frac{X^{2n} - 1}{(X - 1)(X + 1)\Phi_n(X)} \\ &= \frac{X^{2(n-1)} + X^{2(n-2)} + \dots + 1}{X^{n-1} + X^{n-2} + \dots + 1} \\ &= X^{n-1} - X^{n-2} + \dots - X + 1 \\ &= \sum_{i=0}^{n-1} (-X)^i. \end{aligned} \quad (3)$$

The main result will follow from (3) and the following lemma.

**Lemma 1** [Sabia and Tesauri 2009]. *For any integers  $m, b \geq 2$ , any prime divisor of  $\Phi_m(b)$  is either a divisor of  $m$  or is congruent to 1 (mod  $m$ ).*

Suppose that  $n \geq 5$  is prime. By Lemma 1 and (3),

$$\Phi_{2n}(2) = \sum_{i=0}^{n-1} (-2)^i = \frac{(-2)^n - 1}{-3} = \frac{2^n + 1}{3}$$

has prime divisors of  $2n$  or primes congruent to 1 (mod  $2n$ ). The prime divisors of  $2n$  are 2 and  $n$ . Since  $2^n + 1$  is odd and  $2^n + 1 \equiv 3 \pmod{n}$ , neither 2 nor  $n$  divides  $(2^n + 1)/3$ . Therefore,

$$p^*(n) \leq (2^n + 1)/3.$$

### Acknowledgments

The author thanks Dr. Peter Johnson, Jr. for his advice on this project during the Auburn REU in Algebraic and Discrete Mathematics and Dr. David Zureick-Brown for his guidance and encouragement over the past year.

## References

- [Dirichlet 1889] G. L. Dirichlet, *Dirichlet Werke*, G. Reimer, Berlin, 1889. JFM 21.0016.01
- [Heath-Brown 1992] D. R. Heath-Brown, “Zero-free regions for Dirichlet  $L$ -functions, and the least prime in an arithmetic progression”, *Proc. London Math. Soc.* (3) **64**:2 (1992), 265–338. MR 93a:11075 Zbl 0739.11033
- [Linnik 1944a] U. V. Linnik, “On the least prime in an arithmetic progression, I: The basic theorem”, *Rec. Math. [Mat. Sbornik] N.S.* **15**(57) (1944), 139–178. MR 6,260b Zbl 0063.03584
- [Linnik 1944b] U. V. Linnik, “On the least prime in an arithmetic progression, II: The Deuring–Heilbronn phenomenon”, *Rec. Math. [Mat. Sbornik] N.S.* **15**(57) (1944), 347–368. MR 6,260c Zbl 0063.03585
- [Sabia and Tesauri 2009] J. Sabia and S. Tesauri, “The least prime in certain arithmetic progressions”, *Amer. Math. Monthly* **116**:7 (2009), 641–643. MR 2549382 Zbl 1229.11012
- [Thangadurai and Vatwani 2011] R. Thangadurai and A. Vatwani, “The least prime congruent to one modulo  $n$ ”, *Amer. Math. Monthly* **118**:8 (2011), 737–742. MR 2012i:11089 Zbl 1269.11007
- [Xylouris 2009] T. Xylouris, *Über die Linniksche Konstante*, Ph.D. dissertation, Diplomarbeit, Universität Bonn, 2009. arXiv 0906.2749

Received: 2013-11-28

Revised: 2014-03-09

Accepted: 2014-03-20

jmorro2@emory.edu

Emory University, Druid Hills, GA 30306, United States



## Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the *Involve* website.

**Originality.** Submission of a manuscript acknowledges that the manuscript is original and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

**Language.** Articles in *Involve* are usually in English, but articles written in other languages are welcome.

**Required items.** A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

**Format.** Authors are encouraged to use  $\LaTeX$  but submissions in other varieties of  $\TeX$ , and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

**References.** Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of Bib $\TeX$  is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

**Figures.** Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to [graphics@msp.org](mailto: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

2015

vol. 8

no. 2

Enhancing multiple testing: two applications of the probability of correct selection statistic	181
ERIN IRWIN AND JASON WILSON	
On attractors and their basins	195
ALEXANDER ARBIETO AND DAVI OBATA	
Convergence of the maximum zeros of a class of Fibonacci-type polynomials	211
REBECCA GRIDER AND KRISTI KARBER	
Iteration digraphs of a linear function	221
HANNAH ROBERTS	
Numerical integration of rational bubble functions with multiple singularities	233
MICHAEL SCHNEIER	
Finite groups with some weakly $s$ -permutably embedded and weakly $s$ -supplemented subgroups	253
GUO ZHONG, XUANLONG MA, SHIXUN LIN, JIAYI XIA AND JIANXING JIN	
Ordering graphs in a normalized singular value measure	263
CHARLES R. JOHNSON, BRIAN LINS, VICTOR LUO AND SEAN MEEHAN	
More explicit formulas for Bernoulli and Euler numbers	275
FRANCESCA ROMANO	
Crossings of complex line segments	285
SAMULI LEPPÄNEN	
On the $\varepsilon$ -ascent chromatic index of complete graphs	295
JEAN A. BREYTENBACH AND C. M. (KIEKA) MYNHARDT	
Bisection envelopes	307
NOAH FECHTOR-PRADINES	
Degree 14 2-adic fields	329
CHAD AWTRY, NICOLE MILES, JONATHAN MILSTEAD, CHRISTOPHER SHILL AND ERIN STROSNIDER	
Counting set classes with Burnside's lemma	337
JOSHUA CASE, LORI KOBAN AND JORDAN LEGRAND	
Border rank of ternary trilinear forms and the $j$ -invariant	345
DEREK ALLUMS AND JOSEPH M. LANDSBERG	
On the least prime congruent to 1 modulo $n$	357
JACKSON S. MORROW	



1944-4176(2015)8:2;1-5