Springer Proceedings in Mathematics & Statistics

Jan Rychtář
Maya Chhetri
Sat Gupta
Ratnasingham Shivaji *Editors*

Collaborative Mathematics and Statistics Research

Topics from the 9th Annual UNCG Regional Mathematics and Statistics Conference

Springer

Springer Proceedings in Mathematics & Statistics

Volume 109

More information about this series at http://www.springer.com/series/10533

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Jan Rychtář • Maya Chhetri • Sat Gupta
Ratnasingham Shivaji
Editors

Collaborative Mathematics and Statistics Research

Topics from the 9th Annual UNCG Regional Mathematics and Statistics Conference

Springer

Editors
Jan Rychtář
Department of Mathematics and Statistics
University of North Carolina Greensboro
Greensboro, NC, USA

Maya Chhetri
Department of Mathematics and Statistics
University of North Carolina Greensboro
Greensboro, NC, USA

Sat Gupta
Department of Mathematics and Statistics
University of North Carolina Greensboro
Greensboro, NC, USA

Ratnasingham Shivaji
Department of Mathematics and Statistics
University of North Carolina Greensboro
Greensboro, NC, USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
ISBN 978-3-319-11124-7 ISBN 978-3-319-11125-4 (eBook)
DOI 10.1007/978-3-319-11125-4
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014955059

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The annual University of North Carolina Greensboro Regional Mathematics and Statistics Conference (UNCG RMSC) has provided a venue for student researchers to share their work since 2005. The conference is a 1-day event promoting student research in mathematics, statistics, and their applications in various fields. The 2013 conference was held on Saturday, November 2, 2013.

In 2013, the conference was attended by a diverse group of 157 participants from a total of 35 different schools and colleges. Majority of the participants came from UNCG (39), Winthrop University (20), Elon University (13), Bennett College (10), Clemson University (9), Kennesaw State University (7), Clayton State University (6), and NC State University (6). There were a total of 115 student participants which included 3 high-school students, 68 undergraduate students, and 44 graduate students. Many of the participants came from groups underrepresented in STEM disciplines (61 students and 17 faculty were women, 21 participants were African American, 3 were American/Alaskan Native, and 2 were Hispanic).

The students delivered a total of 57 presentations with 29 presentations by undergraduate students, 26 by graduate students, and as a special highlight of the conference, *2 presentations were delivered by high-school students*.

The talks were on various topics of mathematical biology, differential equations, statistics, biostatistics, number theory, algebra, combinatorics, applied mathematics, probability, and computational mathematics. The North Carolina Chapter of the American Statistical Association sponsored the best presentation competition. All presentations were evaluated by a group of faculty volunteers. The following nine students won the *outstanding student presentation award*:

- Graduate student category
  - Tim Antonelli, NC State University
  - Heather Hardeman, Wake Forest University
  - Amanda Traud, NC State University

- Undergraduate student category
  - Ke'Yona Barton and Corbin Smith, Bennett College
  - Eric Kernsfeld, Tufts University
  - Caitlin Ross, UNCG
- High-school student category
  - Manu-Sankara Gargeya, The Early College of Guilford
  - Saumya Goel, Grimsley High School

In addition to the 57 presentations by students, the conference featured two plenary presentations by:

- Simon Tavener, Colorado State University: Evolution of resistance to white pine blister rust in high elevation pines
- Jerry Reiter, Duke University: Protecting data confidentiality in an era without privacy

The conference would not have been possible without the generous support from our sponsors. Funding and support for this conference were provided by the National Science Foundation (grant DMS–1332369), the Department of Mathematics and Statistics, UNCG, and the UNCG College of Arts and Sciences.

All presenters were invited to submit a manuscript to be reviewed for publication in this proceeding. Submitted papers subsequently went through a rigorous refereeing process. The topics covered in this proceeding reflect the topics presented at the conference and the reader will find papers on differential equations, number theory, algebra, combinatorics, probability, statistics, mathematical biology, and computational mathematics.

All papers have a substantial student co-authorship. It is not difficult to look beyond the papers to see the hard work and dedication of many faculty mentors that go well beyond their duties to attract students to research projects in mathematics and statistics. On the other hand, the endless efforts of these excellent students on their research projects are now rewarded by a publication which will encourage them to explore further and deeper. Congratulations to all for this achievement.

We would like to especially highlight the work of two high-school students, Manu-Sankara Gargeya and Saumya Goel. Their manuscripts may not be at the level of other research contributions in this proceeding, but both students and their mentors deserve a warm applause for the work they have done.

We certainly hope the readers will enjoy the manuscripts and will attend the UNCG RMSC conference in the coming years.

Greensboro, NC, USA

Jan Rychtář
Maya Chhetri
Sat Gupta
Ratnasingham Shivaji

Contents

Part 1

Part 2

Part 3

About the Editors

Dr. Jan Rychtář is a professor of mathematics at the Department of Mathematics and Statistics at UNCG. He earned a Ph.D. in 2004 from the University of Alberta, and he joined the UNCG faculty the same year. He works in mathematical biology and game theory. With Professor Mark Broom, he has co-authored a book "Game-Theoretical Models in Biology," and authored or co-authored over 50 papers in peer-reviewed journals. Since 2005, he organizes an annual UNCG Regional Mathematics and Statistics Conference. He has supervised research of over 35 undergraduate students and has served as an Interim Director of the UNCG Office of Undergraduate Research in 2012–2013. He is a councilor for mathematics and computer sciences of the Council of Undergraduate Research.

Dr. Maya Chhetri is a professor of mathematics at the Department of Mathematics and Statistics at UNCG. She received her Ph.D. from Mississippi State University in 1999 and joined UNCG the same year. She works in the area of differential equations and nonlinear analysis. In particular, her research interest is in the study of positive solutions of nonlinear boundary value problems, both ODEs and elliptic PDEs and their applications to other disciplines.

Dr. Sat Gupta received Ph.D. in mathematics from University of Delhi (1977) and Ph.D. in statistics from Colorado State University (1987). He taught at University of Delhi for 6 years, at University of Southern Maine for 18 years, and has been at UNC Greensboro since 2004. He became a Full professor in 1997. His main research area is sampling designs, particularly designs needed for collecting information on sensitive topics where there is a greater likelihood of respondent evasiveness and untruthfulness. He has collaborated with researchers from many fields including biology, marine biology, education, anthropology, psychology, medicine, nursing, and computer science. Some of these collaborative works have been funded by NSF, NIH, and other funding agencies. He is founding editor of the *Journal of Statistical Theory and Practice* (http://www.tandfonline.com/loi/UJSP20) besides serving on the editorial boards of several other journals.

Dr. Ratnasingham Shivaji joined the University of North Carolina at Greensboro (UNCG) as H. Barton Excellence Professor and Head of the Department of Mathematics and Statistics in July 2011. Prior to joining UNCG, he served for 26 years at Mississippi State University (MSU), where he was honored as a W.L. Giles Distinguished Professor. He received his Ph.D. in Mathematics from Heriot-Watt University in Edinburgh, Scotland in 1981 and his B.S. (first class honors) from the University of Sri Lanka in 1977. Shivaji's area of specialization is partial differential equations, and in particular, nonlinear elliptic boundary value problems. His research work has applications in combustion theory, chemical reactor theory, and population dynamics, and has been funded by the National Science Foundation and the Simon's Foundation. To date, Shivaji has authored or co-authored 122 research papers, and served as thesis advisor for 11 Ph.D. graduates.

Part 1

Exploration into the Harmonic Structure of the Tabla

Manu-Sankara B. Gargeya and Promod Pratap

1 Introduction

The Tabla is an Indian percussion instrument that is made up of two main drums, the Tabla (treble drum) and the dagga (bass drum). It originated from the Pakhawaj, a single drum with heads on either side. The original transition from the Pakhawaj to the Tabla is thought to have happened sometime during the eighteenth century. The main components of the Tabla are the pudi (drum head), shell (typically made of rosewood), and gajara (woven goat or camel hide around the drum). The pudi consists of the chat (inner membrane), maidan (outer membrane), and gab (black center). The gab is made from rice paste mixed with iron filings which is baked onto the pudi. The gab serves as the main pitch differentiator of the drum [4].

The different components of the Tabla create a unique harmonic structure for the instrument, first demonstrated experimentally by Raman in [5]. In this study we demonstrate that unlike other percussion instruments, the harmonics of the Tabla are integer multiples of a fundamental.

M.-S.B. Gargeya
The Early College at Guilford, Greensboro, NC 27410, USA
e-mail: gargeyamb2@guilford.edu

P. Pratap (✉)
Department of Physics and Astronomy, The University of North Carolina at Greensboro,
Greensboro, NC 27402, USA
e-mail: prpratap@uncg.edu

J. Rychtář et al. (eds.), *Collaborative Mathematics and Statistics Research*, Springer
Proceedings in Mathematics & Statistics 109, DOI 10.1007/978-3-319-11125-4_1

2 Overview of Harmonic Structure

Harmonic structure is an essential part of any musical instrument in order to determine the range and key of the instrument. The harmonic structure involves the overtones created when an instrument is played at a certain pitch. The displayed overtones allow us to determine the harmonic structure of a given instrument. The harmonic structure is mostly related to the shape and the form of an instrument. Two harmonic structure patterns exist. String instruments and closed instruments (violins, timpani, etc.) demonstrate normal harmonic structure where overtones are found at every integer multiple from the original frequency. Odd harmonic structure, found in instruments with an open end (woodwinds, brass, etc.), exists when overtones are found at odd integer multiple of the original frequency [2]. By determining the harmonic structure pattern of an instrument, more information can be used to model where a membrane or string must be depressed in order to create a particular frequency.

3 Understanding of Frequency and Overtone Examination

For all pitches played by instruments, overtones are produced apart from the fundamental frequency. The overtones represent different octaves of the original note at much lower decibel levels [3,6]. By using the program Audacity, the different overtones can be seen and compared to the fundamental frequency. This comparison is what determines the difference in the harmonic structure.

3.1 Breakdown of the Harmonic Structures of the Piano and Clarinet

The piano, being a string instrument, should demonstrate normal harmonic structure. The clarinet, an open-ended woodwind, should demonstrate odd harmonic structure. The multiplier is found by dividing the frequency by the fundamental frequency (frequency at peak 1) [1]. By playing and recording the middle C for both instruments, harmonic structure could be determined.

The data in Table 1 show that harmonics in the piano occur at approximately every integral multiples of the fundamental, indicating that the harmonic structure for the piano is normal. Similar data for the clarinet (Table 2) show that the harmonics occur at approximately odd multiples of the fundamental, indicating that the harmonic structure of the clarinet is odd.

For the clarinet, multiples were only found at odd integers, see Table 2. Since the multiples are odd, the harmonic structure is also odd.

Table 1 Piano frequency data

Peak #	Peak intensity (Db)	Peak frequency (Hz)	Multiplier
1	−22.7	246	1
2	−23.2	524	2.13
3	−35.3	771	3.13
4	−31.1	1,043	4.13

Table 2 Clarinet frequency data

Peak #	Peak intensity (Db)	Peak frequency (Hz)	Multiplier
1	−11	232	1.00
2	−25.4	706	3.04
3	−28	1,163	5.01
4	−30.7	1,617	6.97

Table 3 Tabla frequency data

Peak #	Peak intensity (Db)	Peak frequency (Hz)	Multiplier
1	−31.3	261	1
2	−27.4	509	1.95
3	−32.1	772	2.96
4	−26.5	1,028	3.94

3.2 Harmonic Structure of the Tabla

The methods used to analyze the harmonic structure of the clarinet and the piano were used to examine the harmonic structure of the Tabla. The note *Ta*, a fundamental note in Tabla, which varies among individual instruments was played and recorded. The *Ta* used in our case was of roughly equivalent frequency to a C natural. Once the *Ta* was recorded, the same multiplier analysis was used as in the previous cases.

In the case of the Tabla, multipliers exist at every frequency demonstrating normal harmonic structure, see Table 3.

Conclusion

Our study was able to successfully identify the harmonic structure of the Tabla. The results of these studies are consistent with the results from Raman [5]. The next step in the analysis of the harmonic structure of the Tabla will be the construction of a mathematical model that will allow us to deduce the harmonics of the instrument from physical parameters such as the radial surface mass density distribution and the radial tension on the membrane.

Acknowledgements Manu-Sankara Gargeya is grateful to Mr. Rajen Sukhadia for the countless hours spent teaching him the art of playing the Tabla.

References

1. Alm JF, Walker JS (2002) Time-frequency analysis of musical instruments. Siam Rev 44(3):457–476
2. Bharucha J, Krumhansl CL (1983) The representation of harmonic structure in music: hierarchies of stability as a function of context. Cognition 13(1):63–102
3. Fletcher H (1934) Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure. J Acoust Soc Am 6:59–69
4. Miśra CL (2007) Playing techniques of tabla: Banaras Gharana. Kanishka Publishers, New Delhi
5. Venkata Raman C (1934) The Indian musical drums. Proc Indian Acad Sci Sect A 1(3):179–188
6. Wolff D (2012) Exploring the physics of sound. Paper written for the PHY 496-02 (individual study), The University of North Carolina at Greensboro

Correlation Between Body Measurements of Different Genders and Races

Saumya Goel and Rahman Tashakkori

1 Introduction

Phidias, a Greek sculptor and mathematician, first discovered ϕ, commonly known today as the Golden Ratio. Phidias studied the phenomenon of ϕ in various Greek sculptures, but Leonardo Da Vinci coined the term "Golden Ratio" by using it in some of his most famous works: "The Last Supper" and "Mona Lisa." In the portrait of Mona Lisa, the wife of an affluent Florence businessman, Da Vinci included numerous examples of the Golden Ratio, as he believed that the Golden Ratio represented an aesthetic bond between humanity and nature.

Human heads vary slightly in size and in different dimensions [5]. In this experiment, the only head dimensions that were considered were head width and head height and dimensions within the facial features. The "Golden Ratio," 1.618, is often considered the most visually appealing ratio in face proportions and other subjects, such as plants and trees. The fig tree is a good example of this ratio. However, there is not sufficient data supporting that the exact "Golden Ratio" is actually common within the human face [10]. Head size does not increase dramatically in small periods of time except in early life and when a head injury occurs [5]. Head height reaches its adult size in both girls and boys at the age of 13. However, head width continues to grow in most people until the age of 18 [2].

S. Goel
Grimsley High School, Greensboro, NC 27408, USA
e-mail: goelsaumya@yahoo.com

R. Tashakkori (✉)
Department of Computer Science, The Appalachian State University, Boone, NC 28608, USA
e-mail: rt@cs.appstate.edu; r.tashakkori@gmail.com

J. Rychtář et al. (eds.), *Collaborative Mathematics and Statistics Research*, Springer Proceedings in Mathematics & Statistics 109, DOI 10.1007/978-3-319-11125-4_2

Height can often accurately be estimated using arm span [1]. Correlation between arm span and height are often stronger among younger subjects; the two measurements are almost interchangeable. As people age, correlation often decreases. This is due to decreasing height in older subjects without significant decrease in arm span [7]. Leg length has also been known to correlate with height measurements, but arm span continues to provide the most reliable estimate [8]. There is strong existing data that humans are taller in the morning and shorter in the evening. This is because pressure on the spine during the day causes loss of fluid content in the intervertebral discs [3]. This was taken into account when measuring height in the experiment as students were asked to participate in the study in the morning and early afternoon.

All of the data gathered in this study was from a summer camp in 2012, Summer Ventures in Science and Mathematics at Appalachian State University. The ratios between arm span and height, and between head width and head height were analyzed to determine possible correlations. Head width is defined as the distance between the widest points on the left and right sides of the face. Head height was defined as the distance between the eye level and the chin. Formal statistical analyses such as t-tests and linear regression were then run on the data for testing the significance of any differences and correlations found. The concept of the Golden Ratio in humans was explored in both genders and all four races.

The outline of the paper is as follows. In Sect. 2, the Golden Ratio is described. Section 3 describes the sample. In Sect. 4, various body characteristics are introduced and in Sect. 5, descriptions of the distribution of these characteristics across races are examined. Statistical analyses and conclusions are discussed in section "Conclusion".

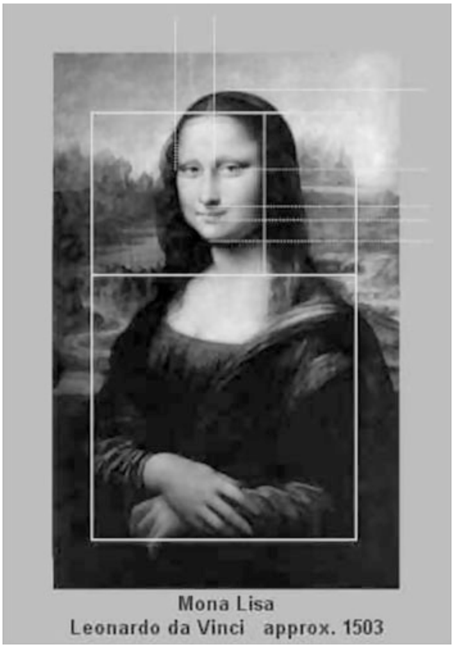
2 Golden Ratio

The distance from the bottom of Mona Lisa's right fingers to the top of her forehead is 1.618 times the distance from the bottom of her right fingers to the base of her neck. The right side of her face is then in golden proportion to the smaller of the original golden rectangle, as shown in Fig. 1.

Other ratios that are considered part of the Golden Ratio are: the distance starting from the base of the neck to the center of the pupil, and the base of the neck to the top of the forehead; the distance from the right side of the face to the right side of top of the nose, and the width of the face at that point; the distance from the bottom of the chin to the bottom of the lips, and the bottom of the chin to the bottom of the nose.

Although the Golden Ratio and Phenomenon of ϕ were invented in the eighteenth century, some of its principles are commonly used today, especially in architecture. The greatest example of Phi is the Parthenon as it uses the Golden Ratio measurements to obtain beauty and equilibrium to its design. Moreover, Notre Dame in Paris contains various Golden Ratio measurements despite its asymmetrical design.

Fig. 1 Mona Lisa



In Asia, the Golden Ratio proportions were employed in the building of the Taj Mahal, one of the Seven Wonders of the World. In Canada, the CN tower is a major example of the Golden Ratio in architecture as the ratio between the observation deck and its total height is exactly 1.618 [6]. Some argue that the phenomenon of Phi in nature is an accident while others sternly believe that it is apparently to add aesthetic appeal.

The phenomenon of ϕ also exists in art. As represented by the "Mona Lisa" by Leonardo Da Vinci, the phenomenon of ϕ is present in other works as well. Many of Da Vinci's famous paintings, including "The Last Supper," contain the Golden Ratio, as in the Renaissance period, Phi was known as the Divine Proportion. Other works such as "The Creation of Adam" by Michelangelo, "The Birth of Venus" by Botticelli, and "The Sacrament of Last Supper" by Salvador Dali accurately reflect the Golden Ratio.

3 Sample

Twenty-eight students from the Visual and Image Processing and the Climate Science classes of the Summer Ventures in Science and Math Program participated in the experiment. Due to the fact that this study was conducted at a high school level summer camp, the available resources were limited, especially with the number and

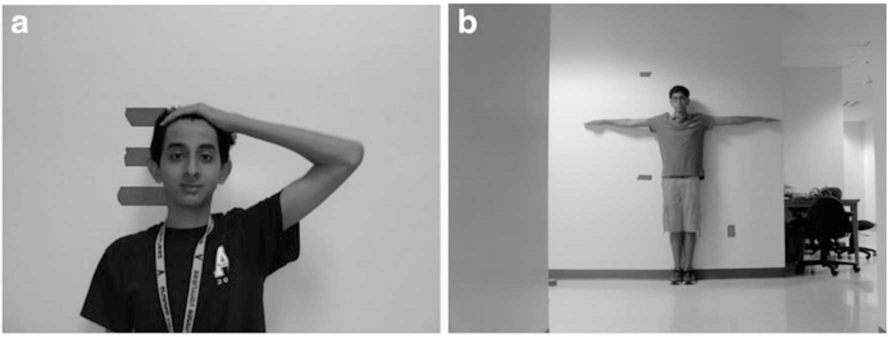


Fig. 2 (**a**) Face shot, (**b**) full body shot

demographics of the subjects. Nevertheless, 14 girls and 14 boys volunteered to participate in the study.

A full-face picture with the hair pulled back and behind the ears and a full-body picture with arms out and feet together was taken for each student. As demonstrated by Fig. 2a, blue tape was used to measure out 10 cm in the face shot. In Fig. 2b, the full body shot, the blue tape represents 1 m. This was done so a scale could be set on ImageJ to determine measurements from photos. ImageJ is a Java-based image-processing program developed at the National Institute of Health (http://imagej.nih.gov/ij/docs/intro.html). With ImageJ's user-written plugin accessibility, it is possible to solve many image processing and analysis problems via different toolboxes. The students' arm span, height, head height, head width, forehead height, and lower face lengths were measured using ImageJ. For the purpose of our experiment, forehead was defined as the area between the hairline and the eye level, the lower face was defined as the distance between the eye level and chin, and head width was measured from the widest point on the face.

People from four different racial backgrounds were present in the sample, but South Asians and Caucasians were dominant in the selected sample. The subjects were separated by race because of different genetic backgrounds. Based on information collected, ratios between different signature measurements can be discovered and used to find defining characteristics of gender and race.

4 Body Characteristics

4.1 Arm Span vs. Height

Figure 3 shows the correlation between arm span and height. Each point represents one subject from the study. The trend line has a slope of 0.7546 and an R^2-value of 0.7799. Since the R^2-value is close to one, it can be concluded that there is a

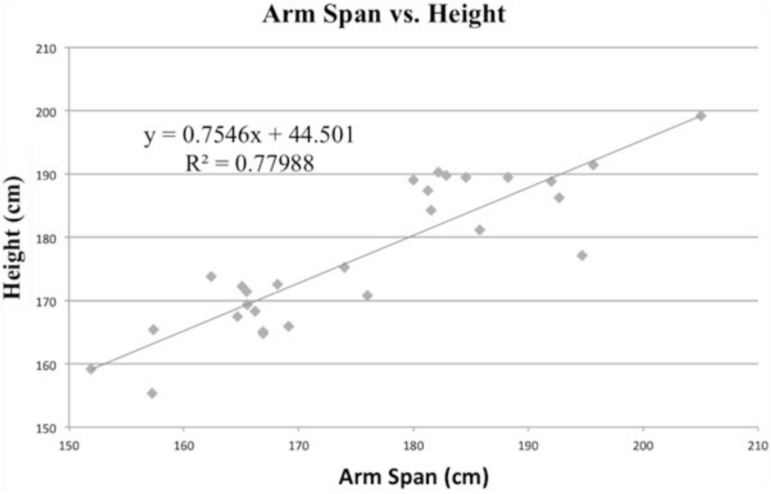


Fig. 3 Arm span vs. height

significant correlation between arm span and height. As seen by our trend line and scatter plot, very little variation is seen around the line of best fit, meaning that the line of best fit matches the data points fairly well. Additionally, a slope *p*-value of 0.000 indicates that the results are statistically significant. It may be noted that the typical *p*-value level for statistical significance is 0.05 or less. Thus, arm span and height indeed do have a significant correlation.

4.2 Head Width vs. Head Height

Figure 4 shows the correlation between head width and head height using data from all participants. The slope of the line of best fit is close to one; however, the R^2-value looks small. In this case also, the slope *p*-value is 0.000, indicating that this correlation is statistically significant as well. The low R^2-value may be due to some outliers towards the lower far right. It is also imperative to keep into consideration that the sample of 28 students for this study was small therefore not providing sufficient reliability.

4.3 Gender Averages

In Fig. 5, the graph compares the male, female, and overall average of arm span, height, head width, head height, forehead height, and lower face measurements. The averages for facial measurements do not differ greatly between the two genders.

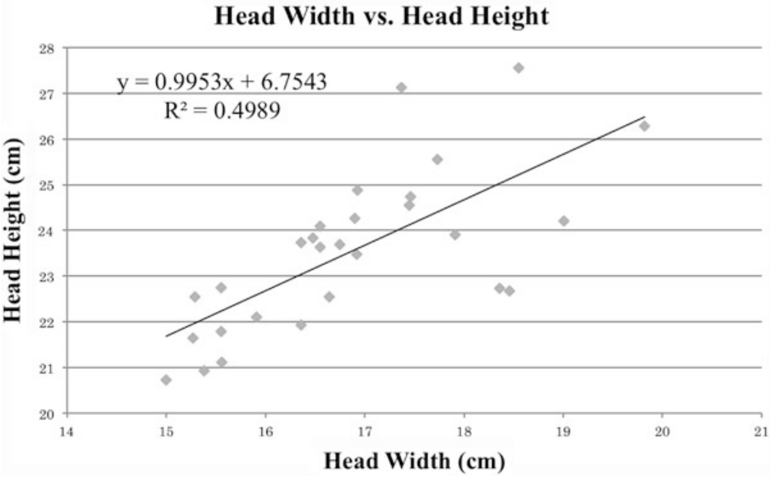


Fig. 4 Head width vs. head height

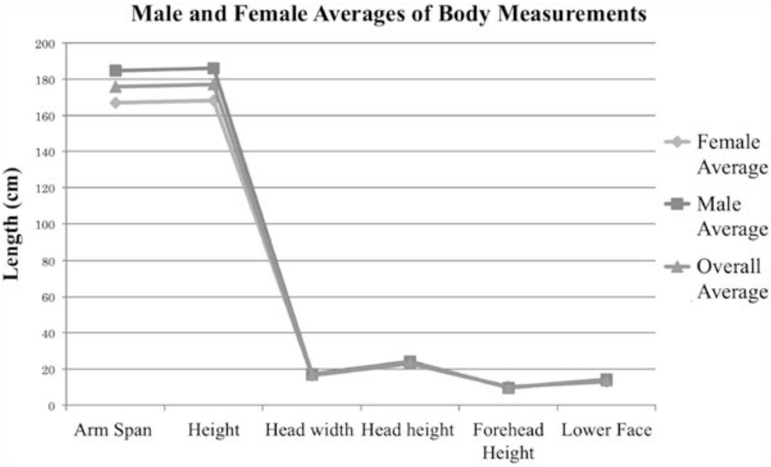


Fig. 5 Gender averages of body measurements

However, the largest difference among male and female averages is arm span and height. The purpose of adding arm span average and height average is to determine whether a correlation between the two exists as hypothesized. For this same reason, the other four averages are added. In essence, the comparison is being made between three pairs, not all six of the measurements. According to the data collected in this experiment, female subjects have a lower average arm span and height than those of the male subjects.

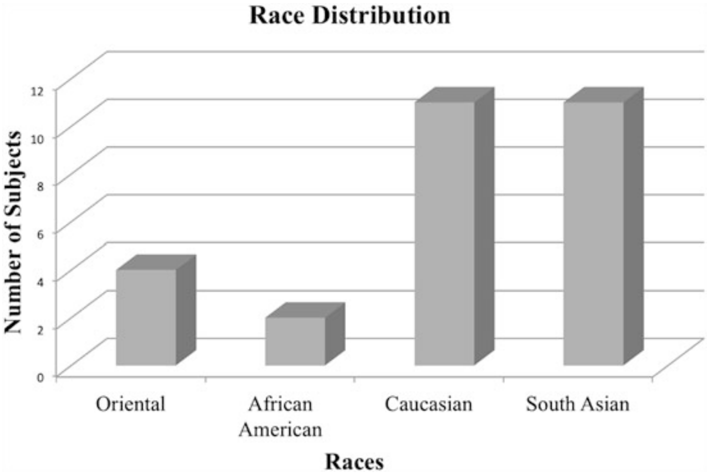


Fig. 6 Race distribution

5 Race Distribution

Figure 6 shows the distribution of participants in the experiment by race. A large majority of the students who were sampled were Caucasian and South Asian. Because this study was conducted at a summer math and science program, the number of South Asians in the study is on expected lines, as historically, South Asians are more interested in math and science as opposed to other subjects [4]. There were fewer Orientals and African Americans among the subjects who participated in this experiment. The term Orientals refers to people from East Asia. This is important to keep in mind when analyzing the data on this experiment because a larger sample size of Orientals and African Americans could have improved the experiment and produced more generalized conclusions.

5.1 Arm Span vs. Height

Figure 7 illustrates the comparison of arm span and height, separated by the four different races present in this study. The data collected in this experiment indicates that African Americans often have larger heights and arm spans. Caucasians had the shortest arm spans and heights, while South Asians had arm spans and heights that were most similar in length. However, as it is evident, the data are heavily skewed because the number of African Americans included in the study was not proportional to the number of Caucasians, Orientals, and South Asians in the study.

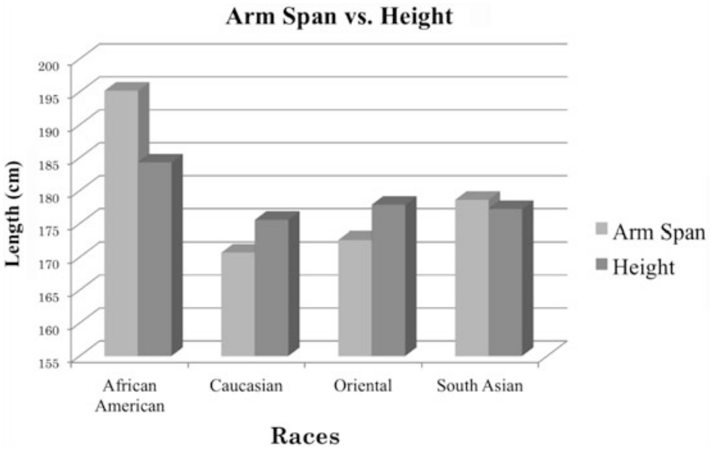


Fig. 7 Arm span vs. height

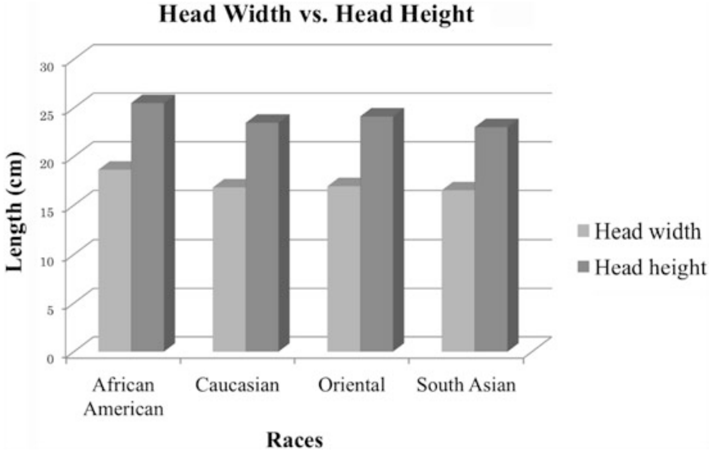


Fig. 8 Head width vs. head height

5.2 *Head Width vs. Head Height*

Figure 8 compares head width with head height separated by race. A consistent trend of larger head heights than head widths was observed in all four racial groups. However, the data in this study shows that African Americans in our study on average had bigger head heights than the other races present in the study. However, not much significance can be attached to this observation since the number of African American participants in the study was very small. It was also observed that the Caucasians in the experiment had a slightly smaller head height. The head

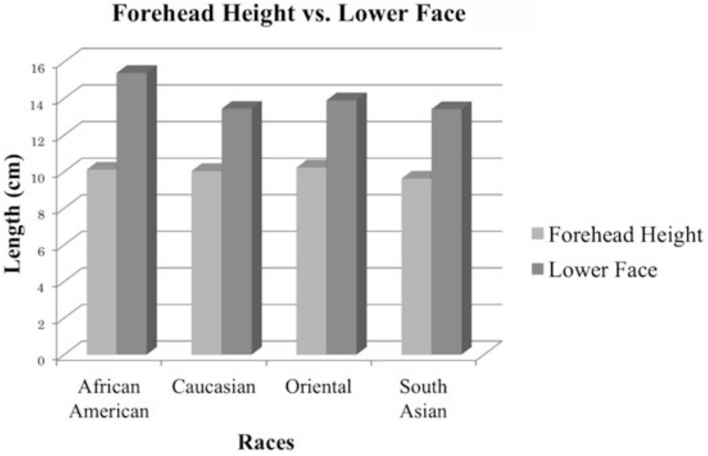


Fig. 9 Forehead height vs. lower face

heights vary from 22 to 25 cm and the head widths vary from 15 to 18 cm. Again, the data are heavily skewed; hence, a general conclusion is difficult to make.

5.3 Forehead Height vs. Lower Face

Figure 9 demonstrates the comparison between forehead height and lower face length, separated by the four races. Using the data collected in this experiment, it can be concluded that African Americans tend to have larger lower faces, while Orientals tend to have larger foreheads.

When analyzing the data, an interesting peculiarity became evident. In Figure H, arm span vs. height, African Americans on average had a greater arm span than height. However, the other races had a greater height than arm span. In the other two graphs, the head height and lower face length were dominant over head width and forehead height for all races. Also, the African Americans stood out in the arm span vs. height graph, as their ratio did not correlate with the ratios of the rest of the races. This was interesting for more extensive study; however, because only two African Americans were involved in this study, it was highly skewed. If each race were represented equally, with a larger sample, the results and ratios would be more reliable.

5.4 Ratios Between Body Measurements

Figure 10 portrays the three different ratios comparing arm span with height, head width with head height, and forehead height with lower face length. In general, the ratio between arm span and height is very close to one in all racial groups.

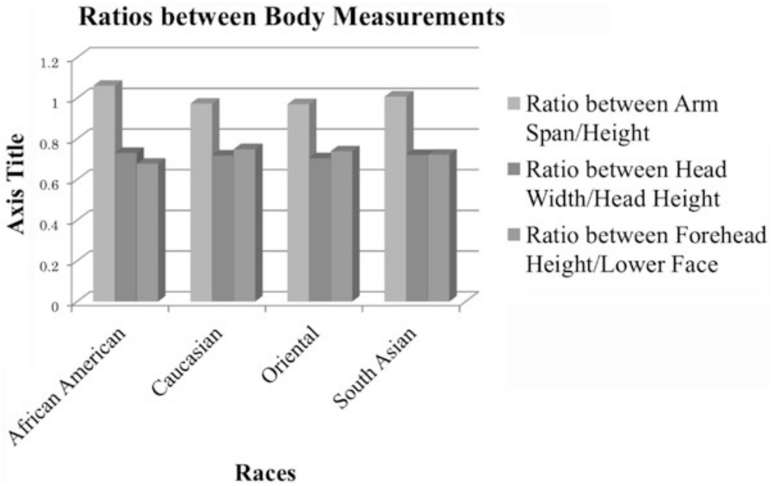


Fig. 10 Ratios between body measurements

However, the ratio between head width and head height and the ratio between forehead height and lower face length is closer to 0.7. Orientals have shorter arm spans in comparison with their height. South Asians have almost the same ratio for head width and head height as forehead height and lower face. It is important to stress that the ratio for African Americans is larger than one but this cannot be statistically proven, as there were only two African American subjects in the study.

Conclusion

The female average of the ratio between the forehead height and the lower face is 0.7897. On the other hand, the average male ratio for the same criteria is 0.6739. A t-test shows that females have a higher average ratio than males (one-sided p-value 0.000). As mentioned above, a p-value of 0.05 or less indicates that the results are statistically significant. The initial goal of this experiment was to test whether it was possible to have someone with the Golden Ratio, i.e., aesthetically as pleasing as Mona Lisa. Unfortunately, there was a large discrepancy between the data set and the Golden Ratio. In this experiment's data set, the closest ratio to the Golden Ratio was 0.8774, which is still only about half of the Golden Ratio. A 95 % confidence interval for the mean ratio in the sample, with all subjects included, turns out to be (0.6975, 0.7661), indicating how unrealistic Mona Lisa's ratio is.

Similar results were also observed from the conclusions of the study done by Ricketts [10], supporting that the Golden Ratio is not common in humans and is therefore only an artistic and architectural concept. Moreover,

(continued)

in another study conducted by Mos et al. [9], similar conclusions were reached as they found that even professional models did not resemble the Golden Ratio. Encompassing participants from different ethnical backgrounds, this study further acknowledges the fact that the Golden Ratio is indeed an ideal proportion for beauty.

In summary, it was concluded that nobody in the data set used in this experiment illustrated a close correlation to the Golden Ratio. However, when women are compared to men, it is obvious that they are more aesthetically pleasing as they have an overall average ratio that is closer to the Golden Ratio than men do.

Acknowledgements The authors wish to offer their sincere gratitude to Professor Sat Gupta of UNCG, Greensboro, NC, for his help with statistical analyses of the data on this study. The authors also acknowledge the comments/suggestions made by the reviewers and the editor which helped improve the presentation of the paper. The authors would like to thank go to the Summer Ventures Program at Appalachian State University and the CS graduate students Ahmad Ghadiri, Sina Tashakkori, and Jack Jordan for their assistance.

References

1. Chhabra SK (2008) Using arm span to derive height: impact of three estimates of height on interpretation of spirometry. Ann Thorac Med 3(3):94–99
2. Farkas LG, Posnick JC, Hreczko TM (1992) Anthropometric growth study of the head. Cleft Palate Craniofac J 29(4):303–308
3. Hutton WC, Malko JA, Fajman WA (2003) Lumbar disc volume measured by mri: effects of bed rest, horizontal exercise, and vertical loading. Aviat Space Environ Med 74(1):73–78
4. Jo L (2012) Asian American college students' mathematics success and the model minority stereotype. Ph.D. thesis, Columbia University
5. Kopperdahl DL, Keaveny TM (1998) Yield strain behavior of trabecular bone. J Biomech 31(7):601–608
6. Livio M (2008) The golden ratio: the story of ϕ, the world's most astonishing number. Random House LLC, New York
7. Manonai J, Khanacharoen A, Theppisai U, Chittacharoen A (2001) Relationship between height and arm span in women of different age groups. J Obstet Gynaecol Res 27(6):325–327
8. Mohanty SP, Suresh Babu S, Sreekumaran Nair N (2001) The use of arm span as a predictor of height: a study of south Indian women. J Orthop Surg 9(1):19–23
9. Moss JP, Linney AD, Lowey MN (1995) The use of three-dimensional techniquesin facial esthetics. In: Seminars in orthodontics. Royal London Hospital Medical School, England, vol 1. Elsevier, pp 94–104. http://www.ncbi.nlm.nih.gov/pubmed/8935048
10. Ricketts RM (1982) The biologic significance of the divine proportion and fibonacci series. Am J Orthod 81(5):351–370

Part 2

The Coupon Collector Problem: The Case of Two Collections

Anda Gadidov and Michael Thomas

1 Introduction

Suppose there are m different cards to complete a certain collection, such as baseball cards or McDonald's Monopoly game pieces. Card of type i occurs independently of the other ones with probability $p_i \geq 0$, $\sum_{i=1}^{m} p_i = 1$. The original question of finding the average number of cards one needs to purchase in order to get a full collection dates back to the early eighteenth century. Assuming equal probabilities the problem was first mentioned in 1709 by DeMoivre in his collection of 26 problems related to games of chance titled *De Mensura Sortis, deu, de Probabilitate Eventuum in Ludis a Casu Fortuito Pendentibus*. In 1938 Kendall and Smith [4] mentioned the problem in relation to checking the randomness of their sampling numbers, Feller [1] presented the question as a type of urn problem, Flajolet et al. [2] used symbolic methods in combinatorial analysis to analyze several related allocation problems. The coupon collector problem is often mentioned in occupancy problems in which balls are thrown independently at a finite or infinite series of boxes. In this context the problem found numerous applications in species sampling problems in ecology, and also in database query optimization.

In 1960 Newman and Schepp [6] generalized the coupon collector's problem to compute the expected number of coupons needed to obtain an arbitrary number of collections under the uniform distribution assumption. Recent generalizations of the problem look into the expected number of coupons collected to complete more than one collection. May [5] analyzes the case when coupons come with different quotas, such as in collecting letters to spell a certain word, using the probability generating function approach. In particular he obtains an expression for the expected value

A. Gadidov (✉) • M. Thomas
Kennesaw State University, 1000 Chastain Rd. #1601, Kennesaw, GA 30144, USA
e-mail: agadidov@kennesaw.edu; mthom130@students.kennesaw.edu

J. Rychtář et al. (eds.), *Collaborative Mathematics and Statistics Research*, Springer Proceedings in Mathematics & Statistics 109, DOI 10.1007/978-3-319-11125-4_3

of the number of coupons needed for n complete collections under the uniform distribution of the coupons (Eq. (7) in [5]).

$$\langle T_{m,n}\rangle = \frac{m^2}{(n-1)!}\int_0^\infty e^{-x}x^n(1-e^{-x}T_n(x))^{m-1}dx \tag{1}$$

where $T_n(x) = 1 + x + \cdots + \frac{x^{n-1}}{(n-1)!}$ is the n-th order Taylor polynomial of the exponential function.

Using the approach in Gadidov and Thomas [3] based on the inclusion–exclusion principle, we obtain the expressions for the probability distribution and the expected value of the number of coupons needed to complete two collections when the coupons have a non-uniform distribution. We also analyze whether the expected value of the number of coupons to complete two collections depends linearly on the number of coupons in the collection using simulations done in R. The paper is organized as follows: the main results are in Theorems 1 and 2 in Sect. 2. In Sect. 3 we build a regression model for the expected number of coupons needed to complete two collections under the uniform distribution of the coupons. The model is based on simulations done in R.

2 Results

The main results are in Theorems 1 and 2. Theorem 1 gives the probability distribution of the random variable X, the number of cards needed to complete two collections, and Theorem 2 gives the expected value of X. If n cards are needed to complete two collections, it means that by the n-th card was drawn a single card was needed to have each type of card at least twice. Define $A_{i,n}$ the event that card of type i was last added when n cards were needed to complete two collections. Then

$$P(X=n) = \sum_{i=1}^{m} P(A_{i,n}) \tag{2}$$

Define X_j the number of cards of type $1 \le j \le m$ in X cards so $X = \sum_{j=1}^{m} X_j$. For a subset of indices $J \subseteq \{1,\ldots,m\}$, denote by $P_J := \sum_{i\in J} p_i$. Also, for $1 \le i \le m$, J_i, $J_{i,0}$ and $J_{i,1}$ will denote subsets of $\{1,\ldots,m\}\setminus\{i\}$ and J_i', $J_{i,0}'$, $J_{i,1}'$ their corresponding complements with respect to $\{1,\ldots,m\}\setminus\{i\}$.

Theorem 1. *For $m \ge 4$ and $n \ge 2m$ we have*

$$P(X=n) = \sum_{i=1}^{m}(n-1)p_i^2(1-p_i)^{n-2} + \sum_{i=1}^{m}\sum_{|J_i|=1}^{m-2}(-1)^{|J_i|}(n-1)p_i^2 P_{J_i'}^{n-2}$$

$$+\sum_{i=1}^{m}\sum_{|J_i|=1}^{m-2}(-1)^{|J_i|}(n-1)\cdots(n-1-|J_i|)p_i^2\prod_{l\in J_i}p_l P_{J_i'}^{n-2-|J_i|}$$

$$+\sum_{i=1}^{m}\sum_{k=2}^{m-2}\sum_{h=1}^{k-1}\sum_{\substack{|J_{i,1}|=h\\|J_{i,0}|=k-h}}\cdots\sum(-1)^k(n-1)\cdots(n-1-h)p_i^2\prod_{l\in J_{i,1}}p_l P_{J_{i,0}'\cap J_{i,1}'}^{n-2-h}\,. \tag{3}$$

The proof of Theorem 1 uses the following generalization of the inclusion–exclusion principle:

Proposition 1. *Let $C_1, C_2, \ldots C_n$ and $D_1, D_2, \ldots D_n$ be sets such that $C_j \cap D_j = \emptyset$ for all $1 \le j \le n$. For $1 < l \le n$ and $1 \le j_1 < j_2 < \cdots < j_l \le n$, denote by $\mathscr{P}_l = \{J \subseteq \{j_1, \ldots, j_l\}\}$ the power set of the set $\{j_1, \ldots, j_l\}$. For $J \in \mathscr{P}_l$, denote by J' the complement of J in $\mathscr{P}_l$. Then the following holds:*

$$P\Big(\bigcup_{j=1}^{n} C_j \cup D_j\Big) = \sum_{j=1}^{n} P(C_j) + \sum_{j=1}^{n} P(D_j)$$
$$+\sum_{l=2}^{n}(-1)^{l+1}\sum_{j_1<j_2<\cdots<j_l}\sum_{k=0}^{l}\sum_{J\in\mathscr{P}_l,|J|=k}P\Big(\bigcap_{j\in J}C_j\cap\bigcap_{i\in J'}D_i\Big)\,. \tag{4}$$

Proof. Let $B_j = C_j \cup D_j$. We know from the inclusion–exclusion principle that

$$P\Big(\bigcup_{j=1}^{n} B_j\Big) = \sum_{j=1}^{n} P(B_j) - \sum_{j<k} P(B_j \cap B_k) + \cdots + (-1)^{l+1}$$
$$\times \sum_{j_1<j_2<\cdots<j_l} P(B_{j_1} \cap B_{j_2} \cap \cdots \cap B_{j_l}) + \cdots + (-1)^n P\Big(\bigcap_{j=1}^{n} B_j\Big)\,.$$

But for every $1 < l \le n$

$$B_{j_1} \cap B_{j_2} \cap \cdots \cap B_{j_l} = \bigcup_{k=0}^{l}\bigcup_{\substack{J\in\mathscr{P}_l\\|J|=k}}\bigcap_{j\in J}C_j \cap \bigcap_{i\in J'}D_i$$

In particular, for $l = 2$ we have $(C_j \cup D_j) \cap (C_i \cup D_i) = C_j \cap C_i \cup C_j \cap D_i \cup D_j \cap C_i \cup D_j \cap D_i$. The result then follows from the fact that $C_j \cap D_j = \emptyset$ for all j. □

We are now ready to give the proof of Theorem 1.

Proof. Let $i \in \{1, 2 \ldots, m\}$. For $n \geq 2m$ we have

$$P(A_{i.n}) = p_i P(X_i = 1, X_j \geq 2, \text{ for all } j \neq i, X = n-1)$$
$$= p_i P(X_i = 1, X = n-1) - p_i P(X_i = 1, X_j \leq 1, \text{ for some } j \neq i, X = n-1) \,. \tag{5}$$

Using the multinomial distribution we get

$$P(X_i = 1, X = n-1) = p_i \sum_{\substack{l_j \geq 0, j \neq i \\ \sum l_j = n-2}} \cdots \sum \frac{(n-1)!}{\prod_{j \neq i} l_j!} \prod_{j \neq i} p_j^{l_j} = p_i (n-1) \big(\sum_{j \neq i} p_j\big)^{n-2}$$
$$= (n-1) p_i (1 - p_i)^{n-2} \,. \tag{6}$$

To evaluate the second probability in (5) we use the inclusion–exclusion formula as in Proposition 1. For $j \neq i$, let $C_j = \{X_i = 1, X_j = 0, X = n-1\}$ and $D_j = \{X_i = 1, X_j = 1, X = n-1\}$. Notice that $C_j \cap D_j = \emptyset$ and

$$P(X_i = 1, X_j \leq 1, \text{ for some } j \neq i, X = n-1) = P(\bigcup_{j \neq i} C_j \cup D_j). \tag{7}$$

Fix $1 \leq i \leq m$ and $J_i \subset \{1, \ldots, m\} \setminus \{i\}, |J_i| = k$. In the following J_i' will denote the complement of J_i in $\{1, \ldots, m\} \setminus \{i\}$. We have

$$P(\bigcap_{j \in J_i} C_j) = P(X_j = 0, j \in J_i, X_i = 1, X = n-1) = \sum_{\substack{i_l \geq 0, l \in J_i' \\ \sum_{l \in J_i'} i_l = n-2}} \cdots \sum p_i \frac{(n-1)!}{\prod_{l \in J_i'} i_l!} \prod_{l \in J_i'} p_l^{i_l}$$
$$= (n-1) p_i \big(\sum_{l \in J_i'} p_l\big)^{n-2} = (n-1) p_i P_{J_i'}^{n-2} \tag{8}$$

$$P(\bigcap_{j \in J_i} D_j) = P(X_j = 1, j \in J, X_i = 1, X = n-1)$$
$$= \sum_{\substack{i_l \geq 0, l \in J' \\ \sum_{l \in J_i'} i_l = n-2-k_1}} \cdots \sum p_i \frac{(n-1)!}{\prod_{l \in J_i'} i_l!} \prod_{j \in J_i} p_j \prod_{l \in J_i'} p_l^{i_l}$$
$$= (n-1) \cdot \cdots \cdot (n-1-k) p_i \prod_{j \in J_i} p_j \big(\sum_{l \in J_i'} p_l\big)^{n-2-k}$$
$$= (n-1) \cdot \cdots \cdot (n-1-k) p_i \prod_{j \in J_i} p_j P_{J_i'}^{n-2-k} \,. \tag{9}$$

Equation (8) gives the probability of having one card of type i and no card of types in J_i in the first $n-1$ cards collected. Equation (9) gives the probability of having one card of type i and of each of the types in J_i in the first $n-1$ cards. Next we will compute the probability of having no card of a given set of types, $J_{i,0}$ and one card of type i and each of the types in $J_{i,1}$, where $J_{i,0}, J_{i,1} \subset \{1,\dots,m\} \setminus \{i\}, |J_{i,0}| = k_0 \geq 1, |J_{i,1}| = k_1 \geq 1, k_0 + k_1 \leq m-2$.

$$\begin{aligned}
P(\bigcap_{h\in J_{i,0}} C_h \cap \bigcap_{j\in J_{i,1}} D_j) &= P(X_h = 0, h \in J_{i,0}, X_j = 1, j \in J_{i,1}, X_i = 1, X = n-1) \\
&= \sum_{\substack{i_l \geq 0, l\in J'_{i,0}\cap J'_{i,1} \\ \sum_{l\in J'_{i,0}\cap J'_{i,1}} i_l = n-2-k_1}} \cdots \sum p_i \frac{(n-1)!}{\prod_{l\in J'_{i,0}\cap J'_{i,1}} i_l!} \prod_{j\in J_{i,1}} p_j \prod_{l\in J'_{i,0}\cap J'_{i,1}} p_l^{i_l} \\
&= (n-1)\cdot\dots\cdot(n-1-k_1) p_i \prod_{j\in J_{i,1}} p_j \Big(\sum_{l\in J'_{i,0}\cap J'_{i,1}} p_l^{i_l}\Big)^{n-2-k_1} \\
&= (n-1)\cdot\dots\cdot(n-1-k_1) p_i \prod_{j\in J_{i,1}} p_j P^{n-2-k_1}_{J'_{i,0}\cap J'_{i,1}} \ . \qquad (10)
\end{aligned}$$

Notice that for $|J_{i,0}|{=}m-1$ or $|J_{i,1}|{=}m-1$, $P(\bigcap_{h\in J_{i,0}} C_h){=}P(\bigcap_{j\in J_{i,1}} D_j) = 0$.

The result then follows from (3) and (5)–(10). □

The next result gives a formula to compute the expected value of the number of cards needed in order to complete two full collections. Let $f(x) = \dfrac{x^{2m}}{1-x}$, $0 < x < 1$, and denote by $f^{(k)}(x)$ the derivative of order k of $f(x)$.

Theorem 2. *The expected value of X is given by*

$$\begin{aligned}
E(X) = &\sum_{i=1}^{m} p_i^2 f^{(2)}(1-p_i) + \sum_{i=1}^{m} \sum_{|J_i|=1}^{m-2} (-1)^{|J_i|} p_i^2 f^{(2)}(P_{J_i'}) \\
&+ \sum_{i=1}^{m} \sum_{|J_i|=1}^{m-2} (-1)^{|J_i|} p_i^2 \prod_{l\in J_i} p_l f^{(|J_i|+2)}(P_{J_i'}) \\
&+ \sum_{i=1}^{m} \sum_{k=2}^{m-2} \sum_{h=1}^{k-1} \sum_{\substack{|J_{i,1}|=h \\ |J_{i,0}|=k-h}} \cdots \sum (-1)^k p_i^2 \prod_{l\in J_{i,1}} p_l f^{(h+2)}(P_{J'_{i,0}\cap J'_{i,1}}) \ . \qquad (11)
\end{aligned}$$

Proof. We have

$$E(X) = \sum_{n=2m}^{\infty} nP(X = n) \ .$$

The result follows from (3) and the fact that for $0 < x < 1$ and positive integer k

$$\sum_{n=2m}^{\infty} n(n-1)\dots(n-k)x^{n-k-1} = \frac{d^{k+1}}{dx^{k+1}}\left(x^{2m}\sum_{n=0}^{\infty} x^n\right) = f^{(k+1)}(x)\ .$$

□

In particular, when the cards in the collection are equally likely, $p_i = 1/m$ for all i and Theorems 1 and 2 become

Corollary 1. *For $m \geq 4$ and $n \geq 2m$ we have*

$$\begin{aligned} P(X=n) &= \frac{n-1}{m}\left(\frac{m-1}{m}\right)^{n-2} + \frac{n-1}{m}\sum_{k=1}^{m-2}(-1)^k\binom{m-1}{k}\left(\frac{m-1-k}{m}\right)^{n-2} \\ &+ \sum_{k=1}^{m-2}(-1)^k\frac{(n-1)!}{(n-2-k)!}\binom{m-1}{k}\left(\frac{1}{m}\right)^{k+1}\left(\frac{m-1-k}{m}\right)^{n-2-k} \\ &+ \sum_{k=2}^{m-2}\sum_{h=1}^{k-1}(-1)^k\frac{(n-1)!}{(n-2-h)!}\cdot\frac{(m-1)!}{h!(k-h)!(m-1-k)!}\left(\frac{1}{m}\right)^{h+1}\left(\frac{m-1-k}{m}\right)^{n-2-h}. \end{aligned} \tag{12}$$

$$\begin{aligned} E(X) &= \frac{1}{m}f^{(2)}\left(\frac{m-1}{m}\right) + \frac{1}{m}\sum_{k=1}^{m-2}(-1)^k\binom{m-1}{k}f^{(2)}\left(\frac{m-1-k}{m}\right) \\ &+ \sum_{k=1}^{m-2}(-1)^k\binom{m-1}{k}\left(\frac{1}{m}\right)^{k+1} f^{(k+2)}\left(\frac{m-1-k}{m}\right) \\ &+ \sum_{k=2}^{m-2}\sum_{h=1}^{k-1}(-1)^k\frac{(n-1)!}{(n-2-h)!}\cdot\frac{(m-1)!}{h!(k-h)!(m-1-k)!}\left(\frac{1}{m}\right)^{h+1} f^{(h+2)}\left(\frac{m-1-k}{m}\right). \end{aligned} \tag{13}$$

Proof. We only need to notice that when cards are equally likely, there are $\binom{m-1}{k}$ subsets $J_i \subset \{1,\dots,m\}\setminus\{i\}$, such that $|J_i| = k$, and for $2 \leq k \leq m-2, 1 \leq h \leq k-1$, there are $\frac{(m-1)!}{h!(k-h)!(m-1-k)!}$ subsets $J_{i,1}, J_{i,0}$ and such that $|J_{i,1}| = h, |J_{i,0}| = k-h$.
The expected value in (13) follows directly from (12) and (11). □

The next theorem gives the probability distribution and the expected value for the cases of collections having two or three cards. We skip its proof as it is based on same methods as the proofs of Theorems 1 and 2.

Theorem 3. *Let $m = 2$ and $0 < p < 1$ be the probability of drawing card of type 1. Then the probability distribution and the expected value of the number of cards needed to complete two collections are given by:*

$$P(X=n)=(n-1)p^2(1-p)^{n-2}+(n-1)(1-p)^2p^{n-2},\ n\geq 4$$

$$E(X)=\frac{2(1-p+2p^3-p^4)}{p(1-p)}\ .$$

Let $m=3$ *and* p_1, p_2, p_3 *the probabilities of the three cards in the collection and let* $g(x)=\dfrac{x^6}{1-x}$. *Then for* $n\geq 6$

$$P(X=n)=(n-1)\sum_{i=1}^{3}p_i^2(1-p_i)^{n-2}-(n-1)\sum_{\substack{i=1\\ j,k\neq i,j<k}}^{3}p_i^{n-2}(p_j^2+p_k^2)$$

$$-(n-1)(n-2)\sum_{\substack{i=1\\ j,k\neq i,j<k}}^{3}p_i^{n-3}(1-p_i)p_jp_k$$

$$E(X)=\sum_{i=1}^{3}p_i^2g^{(2)}(1-p_i)-\sum_{\substack{i=1\\ j,k\neq i,j<k}}^{3}(p_j^2+p_k^2)g^{(2)}(p_i)$$

$$-\sum_{\substack{i=1\\ j,k\neq i,j<k}}^{3}(1-p_i)p_jp_kg^{(3)}(p_i)\ .$$

3 Regression Model

In this section we investigate through simulations the dependence of the expected number of coupons needed to obtain two complete collections on the size of the collection, when coupons are uniformly distributed. It is known that the expected number of coupons needed to complete one collection of size m is

$$E(X)=m\big(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{m}\big)\approx m\ln m \tag{14}$$

We used R software to perform our simulations. The algorithm can be used for any choice of probabilities $p_1, p_2, \ldots, p_m$.

1. For arbitrary probabilities $p_1, p_2, \ldots, p_m$, $\sum_{j=1}^{m} p_j = 1$, partition the interval $[0,1]$ according to $0 = p_0, p_1, p_1+p_2, \ldots, p_1+p_2+\cdots+p_k, \ldots, p_1+p_2+\cdots+p_m=1$.
2. Generate U a uniform random number in $[0,1]$.
3. If $p_0+\cdots+p_{k-1}\leq U < p_0+\cdots+p_k$ for some $k=0,\ldots,m$, then coupon of type k has been pulled.
4. Repeat steps 2 and 3 until each coupon has been obtained twice.

Fig. 1 Linear regression model 1 : $\hat{E}(X) = 6.104\,m - 12.255$

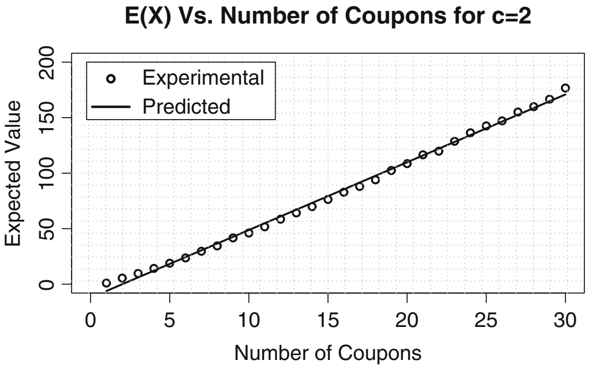


Fig. 2 Residuals model 1

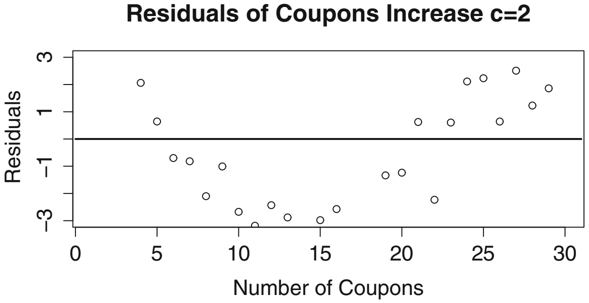
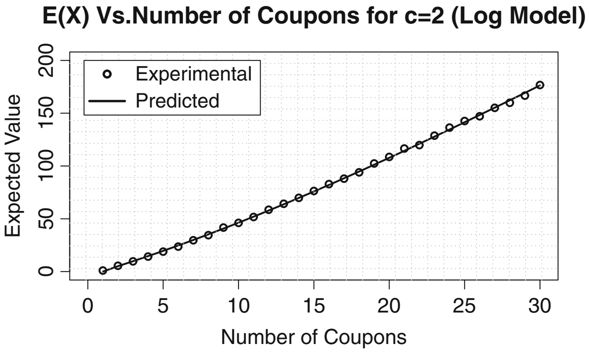


Fig. 3 Linear regression model 2: $\hat{E}(X) = (1.5876m + 4.2203)\ln m$



We ran 500 simulations for each value of m from 1 to 30 and computed the average number of coupons pulled. We built a linear regression model for the expected value of the number of coupons. The first model we built (Fig. 1) returned the equation $\hat{E}(X) = 6.104m - 12.255$ with a value of $R^2 = 0.997$. However, the residual plot displayed a curved pattern (Fig. 2). Based on Eq. (14), $E(X) \approx m \ln m$ for one collection, so we tried a second model after dividing the averages by $\ln m$. The second model returned the equation $\hat{E}(X) = (1.5876m + 4.2203)\ln m$, with an $R^2 = 0.9992$. The model and the residuals are displayed in Figs. 3 and 4, respectively. Based on the residual plot we conclude that the second model is more appropriate.

Fig. 4 Residuals model 2

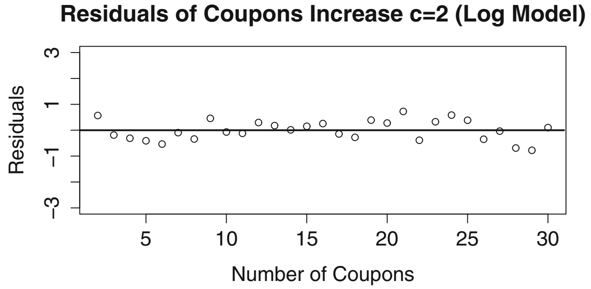


Table 1 Exact, empirical, and predicted expected values

Collection size	Exact mean	Empirical mean	Predicted mean
$m = 2$	5.5	5.52	5.126
$m = 3$	9.64	9.66	9.869
$m = 4$	14.189	14.22	14.654
$m = 5$	19.041	18.91	19.568
$m = 6$	24.134	23.672	24.629

Table 1 gives the exact, empirical, and predicted values (based on the second model) for the expected number of coupons for two collections under the uniform distribution of coupons. The exact values we obtain match the values given by Eq. (1).

Acknowledgements We thank Professor Jan Rychtář from University of North Carolina at Greensboro for making possible our participation at the 9th Annual UNCG Regional Mathematics and Statistics Conference. The conference was sponsored by the National Science Foundation grant DMS-1332369. We also thank the referee for the careful review of the paper, suggestions, and constructive comments.

References

1. Feller W (1968) An introduction to probability theory and its applications, vol I, 3rd edn. Wiley, New York
2. Flajolet P, Gardy D, Thimonier L (1992) Birthday paradox, coupon collectors, caching algorithms and self-organizing search. Discrete Appl Math 39:207–229
3. Gadidov A, Thomas M (2013) The card collector problem. In: Topics from the 8th annual UNCG mathematics and statistics conference. Springer proceedings in mathematics and statistics, vol 64, pp 117–123
4. Kendall MG, Babington Smith B (1938) Randomness and random sampling numbers. J R Stat Soc 101:147–166
5. May R (2008) Coupon collecting with quotas. Electron J Combin 15:2
6. Newman D, Shepp L (1960) The double dixie cup problem. Am Math Mon 67:58–61

Absolute Resolvents and Masses of Irreducible Quintic Polynomials

Chad Awtrey and Christopher R. Shill

1 Introduction

Galois theory stands at the intersection of group theory and field theory. In particular, we can consider field extensions obtained by adjoining to a given base field the roots of monic irreducible polynomials. We are often interested in the arithmetic structure of these field extensions, since this information is related to how the roots of these polynomials interact via the algebraic operations of the base field. Galois theory is important because it associates with each polynomial a group (called its Galois group) that encodes this arithmetic structure. For example, one of the most celebrated results of Galois theory states that an irreducible quintic polynomial is solvable by radicals if and only if its Galois group is a subgroup of the metacyclic group $F_{20} \simeq C_5 \rtimes C_4$.

Therefore, an important problem in computational algebra is to determine the Galois group of an irreducible polynomial defined over a field. Algorithms for accomplishing this task have been in existence for more than a century. Indeed, the original definition of the Galois group implicitly contained a technique for its determination. For an explicit description of this method, see [18, p. 189].

For this paper, we focus on quintic polynomials. Most modern treatments employ a degree 6 resolvent polynomial to determine the Galois group [9, 10]. When this resolvent is factored over the base field, two scenarios can occur. In one case, the resolvent remains irreducible, which indicates the Galois group is either A_5 or S_5,

C. Awtrey (✉)
Elon University, Campus Box 2320, Elon, NC 27244, USA
e-mail: cawtrey@elon.edu

C.R. Shill
Elon University, Campus Box 9017, Elon, NC 27244, USA
e-mail: cshill@elon.edu

J. Rychtář et al. (eds.), *Collaborative Mathematics and Statistics Research*, Springer Proceedings in Mathematics & Statistics 109, DOI 10.1007/978-3-319-11125-4_4

and therefore the original quintic polynomial is not solvable by radicals. In the other case, the degree 6 resolvent factors as a linear times an irreducible quintic, which indicates the Galois group is either cyclic, dihedral, or the previously mentioned metacyclic group of order 20. To identify the Galois group of the original quintic as a transitive subgroup of S_5, more information is needed. To remedy this, Cohen suggests using two additional resolvent polynomials in [9]; this approach is based on Stauduhar's method of relative resolvents [15].

This paper offers two additional methods for computing Galois groups of quintic polynomials. The first is an improvement of the degree 6 resolvent method. The second incorporates what we call the mass of a polynomial, and it is especially useful over finite extensions of the p-adic numbers.

While our results can be applied to any field of characteristic different from 2, we are mostly interested in applications to number fields and p-adic fields. In Sect. 2, we improve upon the results found in [9, 10] by situating the degree 6 resolvent method in the larger framework of what has become known as the absolute resolvent method. In this context, we determine all possible absolute resolvent polynomials as well as their factorizations (which depend on the Galois group of the original quintic). In particular, we show that there is a unique absolute resolvent polynomial of degree 24 the degrees of whose irreducible factors completely determine the Galois group of the original quintic polynomial (i.e., no lower degree resolvent polynomial can accomplish this). Lastly, Sect. 3 introduces a new technique for computing quintic Galois groups based on the notion of the mass of the polynomial. For more examples of the connection between a polynomial's mass and its Galois group see [2–7]. To provide an example of the versatility of our method, we end the paper by examining Galois groups of totally ramified quintic extensions of p-adic fields for primes $p \neq 2, 5$.

2 Absolute Resolvents

Let F be a finite extension of either $\mathbf{Q}$ or $\mathbf{Q}_p$, and let K/F be a quintic extension, defined by the monic irreducible polynomial $f(x)$. Let $\alpha_1, \ldots, \alpha_5$ be the roots of f in some fixed algebraic closure, and let G denote the Galois group of f; i.e., the Galois group of the splitting field of f, or equivalently of the Galois closure of K. Since the elements of G act as permutations on the roots α_i of f, once we fix an ordering of the roots, G can be viewed as a subgroup of S_5 (the symmetric group on 5 letters). Changing the ordering of the roots corresponds to conjugating G in S_5. Since the polynomial f is irreducible, G is a transitive subgroup of S_5; i.e., there is a single orbit for the action of G on the roots α_i of f (each orbit corresponds to an irreducible factor of f).

Therefore, in order to determine the group structure of G, we first identify the conjugacy classes of transitive subgroups of S_5. This information is well known, see [8, 16].

Table 1 Conjugacy classes of nontrivial subgroups of S_5

Name	Transitive?	Size	Generators
C_2	No	2	(12)
C_2^*	No	2	(12)(34)
C_3	No	3	(123)
V_4	No	4	(12)(34), (13)(24)
C_4	No	4	(12)(34), (1324)
V_4^*	No	4	(12), (34)
C_5	Yes	5	(12345)
S_3	No	6	(123), (12)
S_3^*	No	6	(123), (12)(45)
C_3C_2	No	6	(123), (45)
D_4	No	8	(12), (34), (13)(24)
D_5	Yes	10	(12345), (25)(34)
A_4	No	12	(12)(34), (13)(24), (234)
S_3C_2	No	12	(123), (23), (45)
F_{20}	Yes	20	(12345), (25)(34), (2354)
S_4	No	24	(12)(34), (13)(24), (234), (34)
A_5	Yes	60	(12345), (345)
S_5	Yes	120	(12345), (12)

Since we will make use of all conjugacy classes of subgroups of S_5 (not just the transitive subgroups), we include Table 1 for convenience, which gives information on representatives of the 18 conjugacy classes of nontrivial subgroups of S_5. This information can easily be computed with [16].

Most of the group names in the table are standard. For example, C_n represents the cyclic group of order n, D_n the dihedral group of order $2n$, A_n the alternating group on n letters, and V_4 the Klein 4-group (i.e., C_2^2). There are two different conjugacy classes of C_2, V_4, and S_3 in S_5. In each case, we label one group with an asterisk so as to distinguish the conjugacy classes. Note that the only two groups in this table that are not solvable are A_5 and S_5.

2.1 Definition of Resolvents

The usual technique for computing Galois groups involves the notion of absolute resolvent polynomial, which we now define.

Definition 1. Let $T(x_1, \ldots, x_5)$ be a polynomial with integer coefficients. Let H be the stabilizer of T in S_5. That is,

$$H = \{\sigma \in S_5 \colon T(x_{\sigma(1)}, \ldots, x_{\sigma(5)}) = T(x_1, \ldots, x_5)\}.$$

We define the *resolvent polynomial* $R_{f,T}(x)$ of the polynomial $f(x) \in \mathbf{Z}[x]$ by

$$R_{f,T}(x) = \prod_{\sigma \in S_5/H} \left(x - T(\alpha_{\sigma(1)}, \ldots, \alpha_{\sigma(5)})\right),$$

where S_5/H is a complete set of right coset representatives of S_5 modulo H and where $\alpha_1, \ldots, \alpha_5$ are the roots of $f(x)$. By Galois theory, $R_{f,T}(x)$ also has integer coefficients.

2.2 Main Theorem on Resolvents

The main theorem concerning resolvent polynomials is the following. A proof can be found in [14].

Theorem 1. *With the notation of the preceding definition, set $m = [S_5 \colon H] = deg(R_{f,T})$. If $R_{f,T}$ is squarefree, its Galois group (as a subgroup of S_m) is equal to $\phi(G)$, where ϕ is the natural group homomorphism from S_5 to S_m given by the natural right action of S_5 on S_5/H.*

Note that we can always ensure $R_{f,T}$ is squarefree by taking a suitable Tschirnhaus transformation of f [9, p. 324].

As a consequence, this theorem implies that the list of degrees of irreducible factors of $R_{f,T}$ is the same as the length of the orbits of the action of $\phi(G)$ on the set $[1, \ldots, m]$. In particular, $R_{f,T}$ has a root in F if and only if G is conjugate under S_5 to a subgroup of H.

2.3 Example: Discriminant

Perhaps the most well-known example of a resolvent polynomial is the discriminant. Recall that the discriminant of a quintic polynomial $f(x)$ is given by

$$\operatorname{disc}(f) = \prod_{1 \le i < j \le 5} (\alpha_i - \alpha_j)^2,$$

where α_i are the roots of f. In particular, let

$$T = \prod_{1 \le i < j \le 5} (x_i - x_j).$$

It is well known that T is stabilized by A_5 [11, p. 610]. Notice that a complete set of right coset representatives of S_5/A_5 is $\{(1), (12)\}$. Also notice that applying the permutation (12) to the subscripts of T results in $-T$. In this case, we can form the resolvent polynomial as follows:

$$R_{f,T}(x) = \prod_{\sigma \in S_5/A_5} (x - \sigma(T)) = x^2 - T^2 = x^2 - \mathrm{disc}(f).$$

For example, the discriminant of $f(x) = x^5 + a$ is $5^5 a^4$.

2.4 Example: Degree 6 Resolvent

As another example, consider the standard degree 6 resolvent for quintic polynomials [10]. In this case, we let

$$\begin{aligned} T =& x_1^2 x_2 x_5 + x_1^2 x_3 x_4 + x_2^2 x_1 x_3 + x_2^2 x_4 x_5 + x_3^2 x_1 x_5 \\ & + x_3^2 x_2 x_4 + x_4^2 x_1 x_2 + x_4^2 x_3 x_5 + x_5^2 x_1 x_4 + x_5^2 x_2 x_3. \end{aligned}$$

As proven in [10], the stabilizer of T in S_5 is the group F_{20} of order 20, with generators listed in Table 1. A complete set of right coset representatives for S_5/F_{20} is given by $\{(1), (12), (13), (14), (15), (25)\}$. We can form the degree 6 resolvent for $f(x)$ by computing $R_{f,T}$. For example, if $f(x) = x^5 + 2x + 2$, then the degree 6 resolvent of f is $x^6 + 16x^5 + 160x^4 + 1280x^3 + 6400x^2 - 33616x - 283616$ which we computed using [17] to approximate the roots, then rounding the coefficients of the resolvent.

2.5 Determining the Galois Group of a Quintic Polynomial

As stated in Theorem 1, we can use information on how resolvent polynomials factor to determine the Galois group of $f(x)$. Given a transitive subgroup G of S_5, the first step is to determine the factorizations of all possible resolvents arising from a quintic polynomial whose Galois group is G. Theorem 1 shows that this is a purely group-theoretic problem. In particular, the function `resfactors` below will perform such a task. Written for the program GAP [16], the function `resfactors` takes as input a subgroup H of S_5 and a transitive subgroup G of S_5. It computes the permutation representation of G acting on the cosets G/H, and then it outputs the lengths of the orbits of this permutation representation acting on the set $\{1, \ldots, [G : H]\}$. By Theorem 1, the output of the function `resfactors` is precisely the list of degrees of the irreducible factors of $R_{f,T}$ where T is stabilized by H and G is the Galois group of f.

```
resfactors := function(h, g)
  local s5, cosets, index, permrep;
    s5       := SymmetricGroup(5);
```

Table 2 The top row contains the transitive subgroups G of S_5

H/G	C_5	D_5	F_{20}	A_5	S_5
C_2	5,5,5,5,5,5,5,5,5,5,5,5	10,10,10,10,10,10	20,20,20	60	60
C_2^*	5,5,5,5,5,5,5,5,5,5,5,5	5,5,5,5,10,10,10,10	10,10,20,20	30,30	60
C_3	5,5,5,5,5,5,5,5	10,10,10,10	20,20	20,20	40
V_4	5,5,5,5,5,5	5,5,5,5,5,5	10,10,10	15,15	30
C_4	5,5,5,5,5,5	5,5,10,10	5,5,20	30	30
V_4^*	5,5,5,5,5,5	5,5,10,10	10,20	30	30
C_5	1,1,1,1,5,5,5,5	2,2,10,10	4,20	12,12	24
S_3	5,5,5,5	10,10	20	20	20
S_3^*	5,5,5,5	5,5,5,5	10,10	10,10	20
C_3C_2	5,5,5,5	10,10	20	20	20
D_4	5,5,5	5,5,5	5,10	15	15
D_5	1,1,5,5	1,1,5,5	2,10	6,6	12
A_4	5,5	5,5	10	5,5	10
S_3C_2	5,5	5,5	10	10	10
F_{20}	1,5	1,5	1,5	6	6
S_4	5	5	5	5	5
A_5	1,1	1,1	2	1,1	2

The left column contains representatives H of conjugacy classes of subgroups of S_5, as in Table 1. For a particular pair (H, G), the entry in the table gives the output of the function `resfactors(H,G)`. In particular, this list is equivalent to the list of degrees of the irreducible factors of the resolvent polynomial $R(T, f)$ where T is stabilized by H and G is the Galois group of the irreducible quintic polynomial f

```
    cosets  := RightCosets(s5,h);
    index   := Size(cosets);
    permrep := Group(List(GeneratorsOfGroup(g),
               j->Permutation(j, cosets, OnRight)));
  return(List(Orbits(permrep, [1..index]), Size));
end;
```

Table 2 shows for each conjugacy class of subgroups of S_5 the degrees of the irreducible factors of the corresponding resolvent polynomial according to the Galois group of f.

We can use the information in Table 2 to develop algorithms for computing Galois groups of irreducible quintic polynomials. For example, let f be an irreducible quintic polynomial, G the Galois group of f, and g the resolvent for F_{20}. This resolvent is the well-known degree 6 resolvent. We see that:

1. if g factors as a linear times a quintic, then either $G = C_5$, D_5, or F_{20}.
2. if g remains irreducible, then $G = A_5$ or S_5.

The standard procedure for remedying the situation in item (2) is to use the discriminant of f (i.e., the resolvent corresponding to A_5). The discriminant also determines when $G = F_{20}$. To distinguish between C_5 and D_5, we can use the resolvent corresponding to the group C_3C_2, and this is the smallest absolute resolvent that accomplishes that purpose. We point out that Cohen makes use of a different method to distinguish between C_5 and D_5, which is based on Stauduhar's relative resolvent method [15]. See [9] for the details.

As another example, we could use the resolvent corresponding to the group C_2^*, which is a degree 60 polynomial. In this case, all Galois groups are distinguished with this one resolvent. Again letting f denote the quintic polynomial, G its Galois group, and g the degree 60 resolvent corresponding to C_2^*, we have:

1. if g factors as twelve quintics, then $G = C_5$.
2. if g factors as four quintics times four decics, then $G = D_5$.
3. if g factors as two decics times two degree 20 polynomials, then $G = F_{20}$.
4. if g factors as two degree 30 polynomials, then $G = A_5$.
5. if g remains irreducible, then $G = S_5$.

The C_2^* resolvent method mentioned above can be considered an improvement over the degree 6 resolvent method, since it completely determines the Galois group of the polynomial using only one resolvent (and nothing else). However, this method can be improved. If we use the resolvent corresponding to the group C_5, then all Galois groups can still be distinguished with this one resolvent polynomial. However, this only requires factoring a degree 24 resolvent; not a degree 60 polynomial like before. Here is a method to construct this degree 24 resolvent.

A complete set of right coset representatives for S_5/C_5 is:

(1), (45), (34), (132), (125), (35), (23), (135), (152), (12), (1243), (143), (145), (1254), (24), (153), (123), (13), (15), (124), (134), (25), (14), (14)(23).

A form which is stabilized by C_5 is

$$T(x_1, x_2, x_3, x_4, x_5) = x_1x_2^2 + x_2x_3^2 + x_3x_4^2 + x_4x_5^2 + x_5x_1^2.$$

An algorithm for determining quintic Galois groups based on the degree 24 resolvent proceeds as follows. Letting f denote the quintic polynomial, G its Galois group, and g the degree 24 resolvent corresponding to C_5, we have:

1. if g factors as four linears times four quintics, then $G = C_5$.
2. if g factors as two quadratics times two decics, then $G = D_5$.
3. if g factors as one quartic times one degree 20 polynomial, then $G = F_{20}$.
4. if g factors as two dodecics, then $G = A_5$.
5. if g remains irreducible, then $G = S_5$.

For example, consider the Eisenstein polynomial $f(x) = x^5 + 5x + 5$. Forming the degree 24 resolvent polynomial corresponding to the group C_5, we obtain

$$g = x^{24} + 1250x^{21} - 3250x^{20} + \cdots + 2098560333251953125.$$

The resolvent g remains irreducible over $\mathbf{Q}$, indicating the Galois group of f is S_5 in this case. However, g factors as a quartic times a degree 20 polynomial over $\mathbf{Q}_5$ (using [17] for example). Thus the Galois group of f over $\mathbf{Q}_5$ is F_{20}.

3 The Mass of a Polynomial

In the previous section, we situated the standard approach for computing Galois groups of quintic polynomials (the degree 6 resolvent method) into the larger framework of the absolute resolvent method. We completely determined all possible resolvent polynomials along with their factorizations. We then used this information to develop an algorithm to compute Galois groups of quintic polynomials that only relied on factoring a single degree 24 resolvent polynomial.

In this section, we offer a different approach to computing Galois groups of quintic polynomials over p-adic fields. In particular, the aim of this section is to prove the following theorem.

Theorem 2. *Let $p \neq 2, 5$ be a prime number and $F/\mathbf{Q}_p$ a finite extension and let f be its residue degree.*

1. *There are five nonisomorphic totally ramified quintic extensions of F with cyclic Galois group if one of the following conditions holds:*

   a. $p \equiv 1 \pmod 5$*;*
   b. $p \equiv -1 \pmod 5$ *and f is even;*
   c. $p \equiv \pm 2 \pmod 5$ *and* $f \equiv 0 \pmod 4$.

2. *There is a unique totally ramified quintic extension of F whose normal closure has Galois group D_5 if one of the following conditions holds:*

   a. $p \equiv -1 \pmod 5$ *and f is odd;*
   b. $p \equiv \pm 2 \pmod 5$ *and* $f \not\equiv 0 \pmod 4$ *and* $\mathbf{Q}_p(\sqrt{5}) \subset F$.

3. *There is a unique totally ramified quintic extension of F whose normal closure has Galois group F_{20} if* $p \equiv \pm 2 \pmod 5$ *and* $f \not\equiv 0 \pmod 4$ *and* $\mathbf{Q}_p(\sqrt{5}) \not\subset F$.

Notice that this theorem proves the Galois group of an Eisenstein quintic polynomial defined over the p-adic field F depends only on the prime p, the residue degree of F, and whether or not $\sqrt{5} \in F$. Our approach for proving this theorem is to determine the Galois groups of all possible quintic polynomials of a p-adic field simultaneously, rather than focusing on one polynomial at a time.

Toward that end, we fix a prime $p \neq 2, 5$, an algebraic closure $\overline{\mathbf{Q}}_p$ of the p-adic numbers, and a finite extension $F/\mathbf{Q}_p$. We let e be the ramification index of F and

Table 3 Parity and centralizer order for the possible Galois groups of quintic extensions of local fields

G	Parity	$\lvert C_{S_5}(G)\rvert$
C_5	$+$	5
D_5	$+$	1
F_5	$-$	1

let f be its residue degree. Thus $ef = [F : \mathbf{Q}_p]$. For a finite extension K/F, we let K^{gal}/F denote its Galois closure, G the Galois group of K^{gal}/F, and $m(K/F)$ the mass of K/F; that is,

$$m(K/F) = [K : F]/\lvert\mathrm{Aut}(K/F)\rvert,$$

where $\mathrm{Aut}(K/F)$ denotes the automorphism group.

3.1 Two Lemmas

In this subsection, we formulate two technical lemmas and describe how they fit together to yield a proof of Theorem 2. First, we focus on the invariants that distinguish between the possible Galois groups for quintic polynomials. One invariant is the discriminant of the polynomial, mentioned previously. If the Galois group G of the polynomial is a subgroup of A_5, then we say the *parity* of G is $+$. As we saw in the previous section, this occurs precisely when the discriminant of the polynomial is a square in F. Otherwise, we say the parity of G is $-$.

Another invariant we use is the centralizer of G in S_5. This quantity is useful for computing Galois groups since it is isomorphic to the automorphism group of the stem field $F[x]/(f(x))$. It turns out that the parity and centralizer order are enough to distinguish between C_5, D_5, and F_{20} (Table 3). Since Galois groups over local fields are solvable, these three Galois groups are the only cases we need to consider [13, Corollary IV.2.5].

Our remaining lemmas describe how to compute the mass and centralizer order on the field-theoretic side. The first is a standard result for p-adic fields [12, p. 54].

Lemma 1. *Let $F/\mathbf{Q}_p$ be a finite extension and let n be an integer with $p \nmid n$. Let $g = \gcd(p^f - 1, n)$ and let $m = n/g$.*

(a) There are g nonisomorphic totally ramified extensions of F of degree n; each with mass m.

(b) Let ζ be a primitive $(p^f - 1)$-st root of unity and let π be a uniformizer for F. Each totally and tamely ramified extension of F of degree n is isomorphic to an extension that is generated by a root of the polynomial $x^n + \zeta^r\pi$, for some $0 \le r < g$.

Lemma 2. *Let K/F be a totally ramified extension of degree n with $p \nmid n$ and let $g = \gcd(p^f - 1, n)$. Let $G = Gal(K^{gal}/F)$, where K^{gal} is the normal closure of F. Then*

$$g = |C_{S_n}(G)|.$$

Proof. From Galois theory, we know the automorphism group of K/F is isomorphic to the centralizer of G in S_n. Thus the size of $\mathrm{Aut}(K/F)$ is equal to the order of $C_{S_n}(G)$. Using this fact and the definition of the mass of K/F, we have

$$[K : F] = m(K/F) \cdot |\mathrm{Aut}(K/F)| = m(K/F) \cdot |C_{S_n}(G)|.$$

By Lemma 1, we also have

$$[K : F] = m(K/F) \cdot g.$$

These two equations combine to prove the lemma. □

3.2 Proof of Theorem 2

Proof. We know G must be either C_5, D_5, or F_{20}. Let $g = \gcd(p^f - 1, 5)$. Thus g is either 1 or 5. Furthermore, $g = 5$ if and only if $p^f \equiv 1 \pmod 5$, which occurs if either (a) $p \equiv 1 \pmod 5$, (b) $p \equiv -1 \pmod 5$ and f is even, or (c) $p \equiv \pm 2 \pmod 5$ and $f \equiv 0 \pmod 4$. Since $g = |C_{S_5}(G)|$, we see that $G = C_5$ if and only if one of the conditions (a), (b), or (c) occurs; proving part (1). If $g = 1$, then G is either D_5 or F_{20}, depending on whether $\mathrm{disc}(K/F)$ is a square or not, respectively.

Suppose now $g = 1$ and that none of (a), (b), and (c) hold. By Lemma 1, the unique totally ramified quintic extension K/F is generated by a root of the polynomial $x^5 - \pi$ where π is a uniformizer for F. Since K/F is totally ramified, we have

$$\mathrm{disc}(K/F) = \mathrm{disc}(x^5 - \pi) = 5^5\pi^4,$$

which is a square in F if and only if 5 is. Certainly 5 is a square in F if $\mathbf{Q}_p(\sqrt{5}) \subset F$.

Suppose $\mathbf{Q}_p(\sqrt{5}) \not\subset F$, and consider the polynomial $f(x) = x^2 - 5$. Since $p \neq 2, 5$, Hensel's lemma and quadratic reciprocity show that f has a root in F if and only if $p \equiv \pm 1 \pmod 5$. Since we are supposing that (a) and (b) do not hold, it follows that 5 is a square in F if and only if $p \equiv -1 \pmod 5$ and f is odd or $\mathbf{Q}_p(\sqrt{5}) \subset K$. This proves parts (2) and (3). □

We note that with a slight modification to the proof of Theorem 2, the case $p = 2$ can be similarly analyzed. When $p = 5$, the situation is more complicated, but the details can be extracted from [1].

Acknowledgements The authors would like to thank the anonymous reviewer for the careful reading and helpful comments. The authors would also like to thank Elon University for supporting this project through internal grants and the Center for Undergraduate Research in Mathematics for their grant support. This research was supported in part by NSF grant # DMS-1148695.

References

1. Amano S (1971) Eisenstein equations of degree p in a $\mathfrak{p}$-adic field. J Fac Sci Univ Tokyo Sect IA 18:1–21. MR MR0308086 (46 #7201)
2. Awtrey C (2012) Dodecic 3-adic fields. Int J Number Theory 8(4):933–944. MR 2926553
3. Awtrey C (2012) Masses, discriminants, and Galois groups of tame quartic and quintic extensions of local fields. Houston J Math 38(2):397–404. MR 2954644
4. Awtrey C, Shill CR (2013) Galois groups of degree 12 2-adic fields with automorphism group of order 6 and 12. In: Topics from the 8th annual UNCG regional mathematics and statistics conference, vol 64, pp 55–65
5. Awtrey C, Miles N, Milstead J, Shill CR, Strosnider E Degree 14 2-adic fields. Involve (to appear)
6. Awtrey C, Miles N, Shill CR, Strosnider E Computing Galois groups of degree 12 2-adic fields with trivial automorphism group (submitted)
7. Awtrey C, Miles N, Shill CR, Strosnider E Degree 12 2-adic fields with automorphism group of order 4. Rocky Mt J Math (to appear)
8. Butler G, McKay J (1983) The transitive groups of degree up to eleven. Commun Algebra 11(8):863–911. MR 695893 (84f:20005)
9. Cohen H (1993) A course in computational algebraic number theory. Graduate texts in mathematics, vol 138. Springer, Berlin. MR 1228206 (94i:11105)
10. Dummit DS (1991) Solving solvable quintics. Math Comput 57(195):387–401. MR 1079014 (91j:12005)
11. Dummit DS, Foote RM (2004) Abstract algebra, 3rd edn. Wiley, Hoboken. MR 2286236 (2007h:00003)
12. Lang S (1994) Algebraic number theory. Graduate texts in mathematics, vol 110, 2nd edn. Springer, New York. MR 1282723 (95f:11085)
13. Serre J-P (1979) Local fields. Graduate texts in mathematics, vol 67. Springer, New York (Translated from the French by Marvin Jay Greenberg). MR 554237 (82e:12016)
14. Soicher L, McKay J (1985) Computing Galois groups over the rationals. J Number Theory 20(3):273–281. MR MR797178 (87a:12002)
15. Stauduhar RP (1973) The determination of Galois groups. Math Comput 27:981–996. MR 0327712 (48 #6054)
16. The GAP Group (2008) GAP—groups, algorithms, and programming, version 4.4.12
17. The PARI Group (2008) PARI/GP—computational number theory, version 2.3.4. http://pari.math.u-bordeaux.fr/
18. van der Waerden BL (1991) Algebra, vol I. Springer, New York (Based in part on lectures by E. Artin and E. Noether, Translated from the seventh German edition by Fred Blum and John R. Schulenberger). MR 1080172 (91h:00009a)

A Linear Resolvent for Degree 14 Polynomials

Chad Awtrey and Erin Strosnider

1 Introduction

Let p be a prime number. An important problem in computational number theory is to determine the Galois group of an irreducible polynomial f defined over the field of p-adic numbers $\mathbf{Q}_p$. If the degree of f is either equal to p or is not a multiple of p, then it is straightforward to compute the Galois group of f (see, for example, [1, 10]). Otherwise, the situation is more complicated, with no practical general algorithm currently available. However, several researchers have developed ad hoc techniques that depend on both the degree of f and the prime p [2–7, 9–11].

In this paper, we focus on determining the Galois group G when the degree of f is 14 and $p = 7$ (lower degrees have already been treated). Since the elements of G act as permutations on the roots of f, once we fix an ordering on the roots, G can be considered as a subgroup of S_{14}, well defined up to conjugation (different orderings correspond to conjugates of G). Since f is irreducible, G is a transitive subgroup of S_{14}; i.e., there is a single orbit for the action of G on the roots of f (each orbit corresponds to an irreducible factor of f). Therefore our aim is to identify G among the 63 transitive subgroups of S_{14}, following the naming convention that is implemented in [18].

All permutation group computations described in this paper were performed with [18], making extensive use of its transitive group data library. In particular, `GAP` contains all relevant data concerning transitive groups of S_{14} needed for our work. The reliability of `GAP` in this context is supported by the fact that its transitive group data library was authored by Alexander Hulpke, a leading researcher in computational group theory.

C. Awtrey (✉) • E. Strosnider
Elon University, Campus Box 2320, Elon, NC 27244, USA
e-mail: cawtrey@elon.edu; estrosnider@elon.edu

J. Rychtář et al. (eds.), *Collaborative Mathematics and Statistics Research*, Springer Proceedings in Mathematics & Statistics 109, DOI 10.1007/978-3-319-11125-4_5

The remainder of the paper is structured as follows. Section 2 introduces the basic properties of ramification groups to give structural information on possible Galois groups over $\mathbf{Q}_p$. As a consequence of this section, we will show that only 14 of the 63 transitive subgroups of S_{14} are candidates for Galois groups of degree 14 polynomials over $\mathbf{Q}_7$. The goal then becomes to compute enough invariants to uniquely identify the Galois group from among these 14 possibilities. In Sect. 3, we introduce three invariants associated with a polynomial's stem field; namely, the size of its automorphism group, its discriminant, and the Galois groups of its proper, nontrivial subfields. These invariants are enough to distinguish 5 of the 14 possible cases. In the final section, we introduce a linear resolvent polynomial that is able to distinguish the remaining 9 cases. Since the number of isomorphism classes of degree 14 extensions of $\mathbf{Q}_7$ is finite [12, p. 54], it is possible to compute a defining polynomial for each extension and implement our algorithm to compute the polynomial's Galois group. We have carried out this computation, and our results are summarized in Table 1; the final column lists the number of extensions by Galois group.

2 Ramification Groups

The aim of this section is to introduce the basic properties of ramification groups (over general p-adic fields) and use those to deduce structural information about degree 14 extensions of $\mathbf{Q}_7$. A more detailed exposition can be found in [16].

Definition 1. Let $L/\mathbf{Q}_p$ be a Galois extension with Galois group G. Let v be the discrete valuation on L and let $\mathbf{Z}_L$ denote the corresponding discrete valuation ring. For an integer $i \geq -1$, we define the *i-th ramification group* of G to be the following set

$$G_i = \{\sigma \in G : v(\sigma(x) - x) \geq i + 1 \text{ for all } x \in \mathbf{Z}_L\}.$$

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.

Lemma 1. *Let $L/\mathbf{Q}_p$ 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 $\mathbf{Z}_L$ 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 semi-direct 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.*

A proof can be found in [16, Sect. IV].

Specializing to the case where L is the splitting field of an irreducible degree 14 polynomial defined over $\mathbf{Q}_7$, we see that G is a solvable transitive subgroup of S_{14}; of which there are 36. 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 7-group (possibly trivial). Moreover, G_0/G_1 is cyclic of order dividing $7^{[G:G_0]} - 1$. Direct computation on the 36 candidates shows that only 20 are possible Galois groups.

For each of these 20 groups, consider all index 3 subgroups (if there are any); the index 3 subgroups correspond to cubic subfields of L. Now for each such subgroup H, consider the permutation representation of G acting on the cosets of H in G, which is isomorphic to the Galois group of the corresponding cubic subfield. Since all cubic extensions of $\mathbf{Q}_7$ are cyclic (cf. [10]), we can rule out those groups from among the 20 that exhibit an S_3 permutation representation; there are 6 such groups.

Thus, there are 14 possible Galois groups of degree 14 polynomials over $\mathbf{Q}_7$. We identify these groups in the table below using the transitive numbering system in [18]. The second column gives an alternate naming scheme, which is also implemented in `GAP`. Later in the paper, we will reference these groups using only their first column identification.

14T1	C_{14}
14T2	D_7
14T3	D_7C_2
14T4	$2[1/2]F_{42}(7)$
14T5	$F_{21}C_2$
14T7	$F_{42}C_2$
14T8	$C_7 \wr C_2$
14T12	$1/2[D(7)^2]2$
14T13	$[1/2.D(7)^2]2$
14T14	$[7^2 : 3]2$
14T20	$D_7 \wr C_2$
14T23	$[1/6_+.F_{42}^2]2_2$
14T24	$[7^2 : 6]2$
14T32	$[D(7)^2 : 3]2$

3 Stem Field Invariants

As before, let f be a degree 14 polynomial defined over $\mathbf{Q}_7$, and let G be its Galois group. Our aim in this section is to introduce three field-theoretic invariants, related to the stem field of f, that will aid in our computation of G.

First, we consider the stem field of f and its corresponding subgroup H (under the Galois correspondence). Thus H is isomorphic to $G \cap S_{13}$, the point stabilizer of 1 in G. By Galois theory, the automorphism group of the stem field is therefore isomorphic to $N(H)/H$ (where $N(H)$ represents the normalizer of H in G), which is in turn isomorphic to the centralizer of G in S_{14}. In our work, we make use of the size of the automorphism group of the stem field of f, which is equal to the order of the centralizer in S_{14} of G.

Another invariant we employ is related to the discriminant of f. We say the parity of the polynomial f is $+1$ if the discriminant of f is a square in $\mathbf{Q}_7$; otherwise, the parity is -1. On the group theory side, the parity of a polynomial's Galois group is $+1$ if $G \subseteq A_{14}$ and -1 otherwise.

The third invariant we consider is related to the list of the Galois groups of the Galois closures of the proper nontrivial subfields (up to isomorphism) of the stem field of f. We call this the *subfield Galois group* content of f, and we denote it by $sgg(f)$.

Example 1. For example, consider the polynomial $x^{14}+2x^2-2x+3$, which defines the unique unramified degree 14 extension of $\mathbf{Q}_7$. Thus the Galois group G of this polynomial is cyclic of order 14. Since the transitive group notation in [18] lists cyclic groups first, the T-number of G is 14T1. By the fundamental theorem of Galois theory, since G has a unique cyclic subgroup for every divisor of its order, the stem field of f has unique subfields of degrees 2 and 7. These subfields define the unique unramified extensions of $\mathbf{Q}_7$ of their respective degrees, and therefore their Galois groups are also cyclic. Thus the sgg content of f is {2T1, 7T1}.

In general, to compute the sgg content of a polynomial f, we can make use of the complete lists of quadratic and septic 7-adic fields determined in [10] (these lists include defining polynomials along with their Galois groups). For each polynomial in these lists, we can use Panayi's p-adic root-finding algorithm [13,15] to test if the polynomial has a root in the field defined by f. If it does, then this polynomial defines a subfield of the field defined by f. Continuing in this way, it is straightforward to compute the sgg content of f.

The process of employing the sgg content of a polynomial to identify its Galois group is justified by the following result.

Proposition 1. *The sgg content of a polynomial is an invariant of its Galois group (thus it makes sense to speak of the sgg content of a transitive group).*

Proof. Suppose the polynomial f defines an extension L/K of fields, and let G denote the Galois group of f. Let E be the subgroup fixing L/K, arising from the Galois correspondence. The nonisomorphic subfields of L/K correspond to the

Table 1 Invariant data for transitive subgroups of S_{14} that can occur as the Galois group of a degree 14 polynomial defined over $\mathbb{Q}_7$

G	CentOrd	Parity	SGG	F_{91}	Septics	#
14T1	14	-1	2T1,7T1	$7, 14^6$		24
14T2	14	-1	2T1,7T2	$7^7, 14^3$		3
14T8	7	-1	2T1	$14^3, 49$	7T1,7T2	72
14T3	2	-1	2T1,7T2	$7, 14^6$		6
14T5	2	-1	2T1,7T3	$7, 42^2$		24
14T4	2	-1	2T1,7T4	$7, 21^2, 42$		31
14T7	2	-1	2T1,7T4	$7, 42^2$		62
14T12	1	$+1$	2T1	$14^3, 49$	None	8
14T23	1	$+1$	2T1	$42, 49$	None	64
14T13	1	-1	2T1	$14^3, 49$	7T2,7T2	9
14T20	1	-1	2T1	$14^3, 49$	None	16
14T14	1	-1	2T1	42,49	7T3,7T4	93
14T24	1	-1	2T1	42,49	7T4,7T4	114
14T32	1	-1	2T1	42,49	None	128

The column CentOrd gives the order of the group's centralizer in S_{14}, Parity indicates whether the group is even $(+1)$ or not (-1), and SGG gives the *sgg* content of the group. The column F_{91} gives the degrees of the irreducible factors of the linear resolvent F_{91}. When F_{91} has a unique factor of degree 49, Septics gives the Galois groups of all septic subfields of the stem field of this degree 49 factor. The final column gives the number of isomorphism classes of degree 14 extensions of $\mathbb{Q}_7$ whose normal closures have the corresponding Galois group

intermediate subgroups F, up to conjugation, such that $E \leq F \leq G$. Furthermore, 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, every polynomial with Galois group G must have the same subfield content, and this quantity can be determined by a purely group-theoretic computation. □

For each of the 14 possible Galois groups of degree 14 extensions of $\mathbf{Q}_7$, Table 1 shows their respective data for centralizer order, parity, and *sgg* content, with the groups sorted based on their corresponding characteristics. Notice that these three invariants are enough to uniquely identify the five Galois groups 14T1, 14T2, 14T3, 14T5, and 14T8. The final column in the table shows the number of isomorphism classes of degree 14 extensions of $\mathbf{Q}_7$ that have the corresponding group as the Galois group of their normal closure. Note, defining polynomials for these extensions can be computed with a built-in command in [14] (there are a total of 654 such extensions).

To distinguish between the remaining 9 Galois groups, we make use of a linear resolvent (in the sense of [17]).

4 A Linear Resolvent

We begin with a definition of a general resolvent polynomial.

Definition 2. Let $T(x_1, \dots, x_{14})$ be a polynomial with integer coefficients. Let H be the stabilizer of T in S_{14}. That is,

$$H = \left\{\sigma \in S_{14} \colon T(x_{\sigma(1)}, \dots, x_{\sigma(14)}) = T(x_1, \dots, x_{14})\right\}.$$

We define the *resolvent polynomial* $R_{f,T}(x)$ of the polynomial $f(x) \in \mathbf{Z}[x]$ by

$$R_{f,T}(x) = \prod_{\sigma \in S_{14}/H} \left(x - T(r_{\sigma(1)}, \dots, r_{\sigma(14)})\right),$$

where S_{14}/H is a complete set of right coset representatives of S_{14} modulo H and where $r_1, \dots, r_{14}$ are the roots of $f(x)$. By Galois theory, $R_{f,T}(x)$ also has integer coefficients.

The main theorem concerning resolvent polynomials is the following. A proof can be found in [17].

Theorem 1. *With the notation of the preceding definition, set* $m = [S_{14} \colon H] = deg(R_{f,T})$. *If* $R_{f,T}$ *is squarefree, its Galois group (as a subgroup of* S_m*) is equal to* $\phi(G)$*, where* ϕ *is the natural group homomorphism from* S_{14} *to* S_m *given by the natural right action of* S_{14} *on* S_{14}/H*. Note that we can always ensure* R *is squarefree by taking a suitable Tschirnhaus transformation of* f *[8, p. 318].*

As a consequence, this theorem implies that the list of degrees of irreducible factors of $R_{f,T}$ is the same as the length of the orbits of the action of $\phi(G)$ on the set $[1, \dots, m]$. In particular, the Galois group of an irreducible factor of $R_{f,T}$ can be determined by a purely group-theoretic computation.

Our linear resolvent is constructed as follows. Let $T(x_1, \dots, x_{14}) = x_1 + x_2$, which is stabilized by the subgroup $H \simeq S_2 \times S_{12}$ and which is generated by the following three permutations,

$$(1, 2), (3, 4), (3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14).$$

Since $[S_{14} : H] = 91$, the corresponding resolvent polynomial, which we denote by F_{91}, has degree 91, and it can be computed by

$$F_{91}(x) = \prod_{i=1}^{13} \prod_{j=2}^{14} (x - r_i - r_j),$$

where r_i are the roots of the degree 14 polynomial f. However, since T is linear, it can also be computed as a resultant (as in [17]). In particular, let

$$g(x) = \text{Resultant}_y(f(y), f(x+y))/x^{14}.$$

Then $F_{91}(x) = g(\sqrt{x})$.

The list of the irreducible factors of F_{91} is enough to distinguish 4 of the 9 remaining Galois groups (14T4, 14T7, 14T12, and 14T23), as seen in Table 1. For the remaining 5 cases, we note that F_{91} has a unique irreducible factor of degree 49. It turns out that if we consider Galois groups of septic subfields of the stem field of this degree 49 factor, then this information is enough to distinguish between the remaining Galois groups.

Acknowledgements The authors would like to thank Elon University for supporting this project through internal grants and the Center for Undergraduate Research in Mathematics for their grant support. This research was supported in part by NSF grant # DMS-1148695.

References

1. Amano S (1971) Eisenstein equations of degree p in a $\mathfrak{p}$-adic field. J Fac Sci Univ Tokyo Sect IA Math 18:1–21. MR MR0308086 (46 #7201)
2. Awtrey C (2012) Dodecic 3-adic fields. Int J Number Theory 8(4):933–944. MR 2926553
3. Awtrey C (2012) Masses, discriminants, and Galois groups of tame quartic and quintic extensions of local fields. Houst J Math 38(2):397–404. MR 2954644
4. Awtrey C, Miles N, Milstead J, Shill CR, Strosnider E Degree 14 2-adic fields. Involve (to appear)
5. Awtrey C, Miles N, Milstead J, Shill CR, Strosnider E Computing Galois groups of degree 12 2-adic fields with trivial automorphism group (submitted)
6. Awtrey C, Miles N, Milstead J, Shill CR, Strosnider E Degree 12 2-adic fields with automorphism group of order 4. Rocky Mt J Math (to appear)
7. Awtrey C, Shill CR (2013) Galois groups of degree 12 2-adic fields with automorphism group of order 6 and 12. In: Topics from the 8th annual UNCG regional mathematics and statistics conference, vol 64, pp 55–65
8. Cohen H (1993) A course in computational algebraic number theory. Graduate texts in mathematics, vol 138. Springer, Berlin. MR 1228206 (94i:11105)
9. Jones JW, Roberts DP (2004) Nonic 3-adic fields, algorithmic number theory. Lecture notes in computer science, vol 3076. Springer, Berlin, pp 293–308. MR MR2137362 (2006a:11156)
10. Jones JW, Roberts DP (2006) A database of local fields. J Symbolic Comput 41(1):80–97. MR 2194887 (2006k:11230)
11. Jones JW, Roberts DP (2008) Octic 2-adic fields. J Number Theory 128(6):1410–1429. MR MR2419170 (2009d:11163)
12. Lang S (1994) Algebraic number theory, 2nd edn. Graduate texts in mathematics, vol 110. Springer, New York. MR 1282723 (95f:11085)
13. Panayi P (1995) Computation of leopoldt's p-adic regulator. Ph.D. thesis, University of East Anglia
14. PARI Group (2008) The PARI/GP—computational number theory, Version 2.3.4. Available from http://pari.math.u-bordeaux.fr/

15. Pauli S, Roblot X-F (2001) On the computation of all extensions of a p-adic field of a given degree. Math Comp 70(236):1641–1659. MR 1836924 (2002e:11166)
16. Serre J-P (1979) Local fields, graduate texts in mathematics, vol 67 (Translated from the French by Marvin Jay Greenberg). Springer, New York. MR 554237 (82e:12016)
17. Soicher L, McKay J (1985) Computing Galois groups over the rationals. J Number Theory 20(3):273–281. MR MR797178 (87a:12002)
18. The GAP Group (2008) GAP—Groups, algorithms, and programming, Version 4.4.12

Zeros of Partial Sums of the Square of the Riemann Zeta-Function

Kathryn Crosby, Jordan Eliseo, Andrew Ledoan, and David Mazowiecki

2010 *Mathematics Subject Classification.* Primary 11M06; Secondary 11M26

1 Introduction and Statement of Results

The Riemann zeta-function is the analytic function of the complex variable $s = \sigma + it$ that is defined by the absolutely convergent Dirichlet series

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s},$$

K. Crosby
Department of Engineering, Computer Science, Physics, and Mathematics, Oral Roberts University, 7777 South Lewis Avenue, Tulsa, OK 74171, USA
e-mail: kcrosby@oru.edu

J. Eliseo
Department of Mathematics, University of North Carolina at Greensboro, 116 Petty Building, 317 College Avenue, Greensboro, NC 27412, USA
e-mail: jaeliseo@uncg.edu

A. Ledoan (✉)
Department of Mathematics, University of Tennessee at Chattanooga, 415 EMCS Building (Mail Stop 6956), 615 McCallie Avenue, Chattanooga, TN 37403, USA
e-mail: andrew-ledoan@utc.edu

D. Mazowiecki
Department of Mathematics, William Paterson University, 300 Pompton Road, Wayne, NJ 07470, USA
e-mail: mazowieckid@student.wpunj.edu

J. Rychtář et al. (eds.), *Collaborative Mathematics and Statistics Research*, Springer Proceedings in Mathematics & Statistics 109, DOI 10.1007/978-3-319-11125-4_6

valid for $\sigma > 1$. The zeta-function can be continued analytically to a meromorphic function in the s plane with solely a simple pole situated at the point $s = 1$ with residue 1. (See the wonderful expositions of the classical computations by Davenport [4], Ingham [9], and Titchmarsh [18].) Away from the vicinity of the pole, $\zeta(s)$ is unconditionally approximated quite well by arbitrarily short truncations of its Dirichlet series for $\sigma > 1$. This would continue to hold for $\sigma > 1/2$, if the Lindelöf Hypothesis were true. (See Theorem 2.1 in Gonek [5] and Chap. XIII, pp. 328–335, in Titchmarsh [18].)

A great deal is known and conjectured about the distribution of zeros of $\zeta(s)$. However, little is known about the zeros of its partial sums

$$\zeta_X(s) = \sum_{n=1}^{X} \frac{1}{n^s},$$

where the length of truncation X is a real number greater than or equal to 2. Exceptions are the works of Montgomery [11], Turán [19–22], and Voronin [24], with numerical studies carried out by Spira [16, 17] and, more recently, by Borwein et al. [3].

From the absolute convergence of the Dirichlet series defining $\zeta(s)$, a quick calculation confirms that, even for X not very large, $\zeta_X(s)$ provides a rather good approximation to $\zeta(s)$ with a remainder which is $o(1)$ as $X \to \infty$. Motivated in part by the approximate functional equation for $\zeta(s)$ due to Hardy and Littlewood [7, 8], Gonek and one of the authors [6] studied the distribution of zeros of $\zeta_X(s)$. Theorem 1 in [6] sums up a number of known results in the literature on the zeros of $\zeta_X(s)$. We summarize the theorem here for convenience.

Theorem 1. *The following results on the distribution of zeros of $\zeta_X(s)$ are known.*

(a) (Borwein et al. [3].) *Let X be a real number greater than or equal to 2. Then the zeros of $\zeta_X(s)$ all lie in the strip $\alpha < \sigma < \beta$, where the real numbers α and β are the unique solutions of*

$$1 + 2^{-\sigma} + \cdots + (X-1)^{-\sigma} = X^{-\sigma}$$

and

$$2^{-\sigma} + 3^{-\sigma} + \cdots + X^{-\sigma} = 1,$$

respectively. In particular, $\alpha > -X$ and $\beta < 1.72865$.

(b) (Montgomery and Vaughan [12].) *There exists a real number X_0 such that, when $X \geq X_0$, $\zeta_X(s)$ has no zeros for*

$$\sigma \geq 1 + \left(\frac{4}{\pi} - 1\right) \frac{\log\log X}{\log X}.$$

(c) (Montgomery [11].) *For any constant C satisfying the inequalities* $0 < C < 4/\pi - 1$, *there exists a real number X_0 depending on C only such that, when $X \geq X_0$, $\zeta_X(s)$ has zeros for*

$$\sigma > 1 + \frac{C \log\log X}{\log X}.$$

In [6], Gonek and one of the authors extended the above investigations. We can summarize, in particular, Theorem 2 in [6] as follows.

Theorem 2 (Gonek and Ledoan [6]). *Let X and T be real numbers greater than or equal to* 2. *Denote by $N_{\zeta_X}(T)$ the number of zeros of $\zeta_X(s)$ with ordinates in the interval* $[0, T]$, *with*

$$N_{\zeta_X}(T) = \lim_{\epsilon \to 0+} N_{\zeta_X}(T + \epsilon)$$

when T coincides with the ordinate of a zero. Then

$$\left| N_{\zeta_X}(T) - \frac{T}{2\pi} \log[X] \right| < \frac{X}{2},$$

where $[X]$ denotes the greatest integer less than or equal to X.

In relation to these works and motivated in part by the approximate functional equation for $\zeta^2(s)$ due to Hardy and Littlewood [8], our object in this paper is to investigate the distribution of zeros of the partial sums

$$\zeta_X^2(s) = \sum_{n=1}^{X} \frac{d(n)}{n^s}, \tag{1}$$

where $d(n)$ denotes the number of divisors of n. The partial sums $\zeta_X^2(s)$ are intimately connected with $\zeta^2(s)$. Since $\zeta^2(s)$ does not converge for $\sigma \leq 1$, it is difficult to determine what possible relationship might exist between the zeros of $\zeta^2(s)$ and $\zeta_X^2(s)$ for $\sigma \leq 1$.

In this paper, we first derive the zero-free regions of $\zeta_X^2(s)$. We are ready to state our first two main results.

Theorem 3. *There exists a real number β which depends on X only such that $\zeta_X^2(s)$ has no zeros for $\sigma \geq \beta$. In particular,*

$$\beta = \frac{1}{\log 2}\left[1 - \left(\frac{\pi^2}{6} - \frac{5}{4}\right) X^{(2/(\log 3 - 1))(\log X + 1)}\right] + 1.$$

Theorem 4. *There exists a real number α which depends on X only such that $\zeta_X^2(s)$ has no zeros for $\sigma \leq \alpha$. In particular,*

$$\alpha = -\frac{2(\log X + 1)}{\log 3 - 1} X \log X.$$

We then use Theorems 3 and 4 to establish an approximate formula for the number of zeros up to a given height T. Our third main result may be stated as follows.

Theorem 5. *Let X and T be real numbers greater than or equal to* 2. *Denote by $N_{\zeta_X^2}(T)$ the number of zeros of $\zeta_X^2(s)$ with ordinates in the interval* $[0, T]$, *with*

$$N_{\zeta_X^2}(T) = \lim_{\epsilon \to 0+} N_{\zeta_X^2}(T + \epsilon)$$

when T coincides with the ordinate of a zero. Then

$$N_{\zeta_X^2}(T) = \frac{T}{2\pi} \log X + O(X).$$

2 Proof of Theorem 3

In this section, we show that the magnitude of $\zeta_X^2(s)$ is greater than zero for $\sigma \geq \beta$. Toward this end, we extract the first term from the sum in (1), take absolute values, and apply the triangle inequality twice to obtain

$$|\zeta_X^2(s)| \geq 1 - \left| \sum_{n=2}^{X} \frac{d(n)}{n^s} \right| \geq 1 - \sum_{n=2}^{X} \frac{d(n)}{n^\sigma}. \tag{2}$$

In order to determine a real number β such that the quantity on the far right-hand side of (2) is strictly greater than zero for $\sigma \geq \beta$, we will need to make use of an upper bound for $d(n)$. The papers by Nicolas and Robin [13] and Usol'cev [23] contain more general bounds for $d(n)$ from which it follows that, for all integers $n \geq 1$,

$$\log d(n) < \frac{2 \log n}{\log \log n}. \tag{3}$$

(See, also, the handbook by Sándor et al. [15], Chapter II, Section II.3, p. 41.)

We recall that, for all real numbers $x \geq 1$,

$$\log x \geq \frac{x - 1}{x + 1},$$

with equality only when $x = 1$. Hence, for all real numbers $x \geq e$,

$$\log\log x \geq \frac{\log x - 1}{\log x + 1}.$$

It follows that, for all integers n in the range $3 \leq n \leq X$,

$$\begin{aligned} d(n) &< e^{(2\log n)(\log n+1)/(\log n-1)} \\ &\leq e^{(2\log X)(\log X+1)/(\log 3-1)} \\ &\leq X^{(2/(\log 3-1))(\log X+1)}. \end{aligned} \tag{4}$$

Hence, the quantity on the far right-hand side of (2) must satisfy

$$\begin{aligned} 1 - \sum_{n=2}^{X} \frac{d(n)}{n^\sigma} &> 1 - \frac{1}{2^{\sigma-1}} - X^{(2/(\log 3-1))(\log X+1)} \sum_{n=3}^{X} \frac{1}{n^\sigma} \\ &> 1 - \frac{1}{2^{\sigma-1}} - X^{(2/(\log 3-1))(\log X+1)} \sum_{n=3}^{\infty} \frac{1}{n^\sigma}. \end{aligned} \tag{5}$$

In view of the quantities on the far right-hand sides of (2) and (5), it will be sufficient to show that

$$\frac{1}{2^{\sigma-1}} + X^{(2/(\log 3-1))(\log X+1)} \sum_{n=3}^{\infty} \frac{1}{n^\sigma} < 1. \tag{6}$$

Since $\sigma \geq \beta$ and $n \geq 3$,

$$n^\sigma \geq n^\beta = n^{\beta-2}n^2 \geq 3^{\beta-2}n^2.$$

From this and the discovery due to Euler that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6},$$

we see that the sum in (6) satisfies

$$\sum_{n=3}^{\infty} \frac{1}{n^\sigma} \leq \sum_{n=3}^{\infty} \frac{1}{n^\beta} \leq \frac{1}{3^{\beta-2}} \sum_{n=3}^{\infty} \frac{1}{n^2} = \frac{1}{3^{\beta-2}} \left(\frac{\pi^2}{6} - \frac{5}{4} \right), \tag{7}$$

which is substantially the best possible. Furthermore, since $\sigma \geq \beta$,

$$1-\sum_{n=2}^{X}\frac{d(n)}{n^{\sigma}} > 1-\frac{1}{2^{\beta-1}}-\frac{1}{3^{\beta-2}}\left(\frac{\pi^2}{6}-\frac{5}{4}\right)X^{(2/(\log 3-1))(\log X+1)}$$
$$> 1-\frac{1}{2^{\beta-1}}\left[1+\left(\frac{\pi^2}{6}-\frac{5}{4}\right)X^{(2/(\log 3-1))(\log X+1)}\right].$$

In order for (6) to hold true, we must have that

$$\frac{1}{2^{\beta-1}}\left[1+\left(\frac{\pi^2}{6}-\frac{5}{4}\right)X^{(2/(\log 3-1))(\log X+1)}\right] < 1.$$

Solving for β in the last inequality above, we find that β is a real number that must satisfy

$$\beta > \frac{1}{\log 2}\left[1+\left(\frac{\pi^2}{6}-\frac{5}{4}\right)X^{(2/(\log 3-1))(\log X+1)}\right]+1, \tag{8}$$

from which we obtain an admissible value for β.

By virtue of estimates (2) through (8), we deduce that

$$\sum_{n=2}^{\infty}\frac{d(n)}{n^{\beta}} \leq \frac{1}{2^{\beta-1}}\left[1+\left(\frac{\pi^2}{6}-\frac{5}{4}\right)X^{(2/(\log 3-1))(\log X+1)}\right].$$

Hence, for $\sigma \geq \beta$,

$$\left|\sum_{n=2}^{X}\frac{d(n)}{n^{s}}\right| \leq \sum_{n=2}^{X}\frac{d(n)}{n^{\beta}} < 1, \tag{9}$$

and it follows that

$$|\zeta_X^2(s)| \geq 1-\left|\sum_{n=2}^{X}\frac{d(n)}{n^{s}}\right| > 0.$$

Consequently, $|\zeta_X^2(s)| > 0$ for $\sigma \geq \beta$. Hence, $\zeta_X^2(s)$ has no zeros for $\sigma \geq \beta$. This completes the proof of Theorem 3.

3 Proof of Theorem 4

In this section, we show that the magnitude of $\zeta_X^2(s)$ is greater than zero for $\sigma \leq \alpha$. Proceeding to the proof of the theorem, we extract the last term from the sum in (1), take absolute values, and apply the triangle inequality twice to obtain

$$|\zeta_X^2(s)| \geq \frac{d(X)}{X^\sigma} - \left|\sum_{n=1}^{X-1} \frac{d(n)}{n^s}\right| \geq \frac{d(X)}{X^\sigma} - \sum_{n=1}^{X-1} \frac{d(n)}{n^\sigma}.$$

It will be sufficient to obtain a real number α such that, for $\sigma \leq \alpha$,

$$\frac{2}{X^\sigma} > \sum_{n=1}^{X-1} \frac{d(n)}{n^\sigma}. \tag{10}$$

By (3), for any real number $\sigma \leq \alpha$,

$$\begin{aligned}\frac{1}{X^\sigma} &> \frac{1}{X^{\sigma-\alpha}}\left(1 + \frac{1}{2^{\alpha-1}} + \sum_{n=3}^{X-1} \frac{1}{n^\alpha} e^{2\log n/\log\log n}\right)\\ &= \frac{1}{X^{\sigma-\alpha}} + \frac{1}{X^{\sigma-\alpha}2^{\alpha-1}} + \sum_{n=3}^{X-1} \frac{1}{X^{\sigma-\alpha}n^\alpha} e^{2\log n/\log\log n}.\end{aligned}$$

Since $n < X$,

$$\begin{aligned}\frac{1}{X^\sigma} &> 1 + \frac{1}{2^{\alpha-1}} + \sum_{n=3}^{X-1} \frac{1}{n^{\sigma-\alpha}n^\alpha} e^{2\log n/\log\log n}\\ &= 1 + \frac{1}{2^{\alpha-1}} + \sum_{n=3}^{X-1} \frac{1}{n^\sigma} e^{2\log n/\log\log n}\\ &> 1 + \sum_{n=3}^{X-1} \frac{1}{n^\sigma} e^{2\log n/\log\log n}.\end{aligned}$$

It suffices to obtain a real number α such that

$$\frac{1}{X^\alpha} > 1 + \sum_{n=3}^{X-1} \frac{1}{X^\sigma} e^{2\log n/\log\log n}. \tag{11}$$

Since $3 \leq n \leq X-1 < X$, we have $n^\alpha > X^\alpha$. It follows that

$$\frac{1}{X^\alpha} \sum_{n=3}^{X-1} e^{2\log n/\log\log n} > \sum_{n=3}^{X-1} \frac{1}{n^\alpha} e^{2\log n/\log\log n}.$$

It should be observed, however, that it suffices to have

$$\frac{1}{X^\alpha} > 1 + \frac{1}{X^\alpha}\sum_{n=3}^{X-1} e^{2\log n/\log\log n},$$

which holds for any fixed negative real number α and for all X sufficiently large in terms of α. We may take any fixed negative real number α depending on X only for which (11) holds. By (4),

$$\sum_{n=3}^{X-1}\frac{1}{n^\alpha}e^{2\log n/\log\log n} < X^{(2/(\log 3-1))(\log X+1)}\sum_{n=3}^{X-1}\frac{1}{n^\alpha}. \tag{12}$$

It remains to treat the sum on the right-hand side of (12).

When $\alpha < 0$, Cauchy's integral test yields

$$\sum_{n=3}^{X-1}\frac{1}{n^\alpha} \le \frac{1}{(X-1)^\alpha} + \int_3^{X-1}\frac{1}{y^\alpha}\,dy < \frac{X-\alpha}{(1-\alpha)(X-1)^\alpha}.$$

From this and (12), we see that (11) follows directly from

$$\frac{1}{X^\alpha} > X^{(2/(\log 3-1))(\log X+1)}\frac{X-\alpha}{(1-\alpha)(X-1)^\alpha}.$$

Solving for α in the last inequality, we obtain an admissible value for α. In particular,

$$\alpha = -\frac{2(\log X+1)}{\log 3-1}X\log X.$$

Hence, we conclude that $|\zeta_X^2(s)| > 0$ for $\sigma \le \alpha$. Therefore, $\zeta_X^2(s)$ has no zeros for $\sigma \le \alpha$. This finishes the proof of Theorem 4.

4 Proof of Theorem 5

We now take up the proof of the approximate formula for $N_{\zeta_X^2}(T)$. It will be simpler, from purely a mathematical standpoint, to assume that T, which we suppose to be large, does not coincide with the ordinate of a zero of $\zeta_X^2(s)$.

By the principle of the argument (see Ahlfors [1], Theorem 18, p. 152) and Theorems 3 and 4,

$$2\pi N_{\zeta_X^2}(T) = \triangle_R \arg\zeta_X^2(s),$$

where R denotes the rectangle in the s plane with vertices α, β, $\beta+iT$, and $\alpha+iT$, described in the positive sense and $\triangle_R$ denotes the change in the argument of $\zeta_X^2(s)$ as s traverses R in the positive sense.

Here, we note that the change in the argument of $\zeta_X^2(s)$ as s describes the top edge of R is $-\triangle_{[\alpha,\beta]}\arg\zeta_X^2(\sigma+iT)$, the minus sign coming from the fact that along R this side is traced in the negative sense. In the same way, the change in the argument of $\zeta_X^2(s)$ as s describes the left edge of R is $-\triangle_{[0,T]}\arg\zeta_X^2(\alpha+it)$.

Hence,

$$\begin{aligned}2\pi N_{\zeta_X^2}(T) &= \triangle_{[\alpha,\beta]}\arg\zeta_X^2(\sigma)+\triangle_{[0,T]}\arg\zeta_X^2(\beta+it)\\ &\quad -\triangle_{[\alpha,\beta]}\arg\zeta_X^2(\sigma+iT)-\triangle_{[0,T]}\arg\zeta_X^2(\alpha+it),\end{aligned}\tag{13}$$

and we distinguish four cases.

As s describes the base of R, there is no change in $\arg\zeta_X^2(s)$. Since $\zeta_X^2(s)$ is real and nowhere zero on $[\alpha,\beta]$, we have

$$\triangle_{[\alpha,\beta]}\arg\zeta_X^2(\sigma)=0.\tag{14}$$

As s describes the right edge of R, it follows from (9) that

$$|\zeta_X^2(s)-1|<1.$$

Hence, for $0\le t\le T$, we have $\arg\zeta_X^2(\beta)=0$ and $\Re(\zeta_X^2(\beta+it))>0$, so that

$$\triangle_{[0,T]}\arg\zeta_X^2(\beta+it)=O(1).\tag{15}$$

Along the top edge of R, we estimate the change in $\arg\zeta_X^2(\sigma+iT)$ by first decomposing $\zeta_X^2(\sigma+iT)$ into the sum of its real and imaginary parts. These are denoted by $\Re(\zeta_X^2(\sigma+iT))$ and $\Im(\zeta_X^2(\sigma+iT))$, respectively. Hence,

$$\begin{aligned}\zeta_X^2(\sigma+iT) &= \sum_{n=1}^{X} d(n)e^{-(\sigma+iT)\log n}\\ &= \sum_{n=1}^{X}\frac{d(n)}{n^\sigma}[\cos(T\log n)-i\sin(T\log n)]\\ &= \Re(\zeta_X^2(\sigma+iT))+i\Im(\zeta_X^2(\sigma+iT)),\end{aligned}$$

where

$$\Re(\zeta_X^2(\sigma+iT))=\sum_{n=1}^{X}\frac{d(n)}{n^\sigma}\cos(T\log n)$$

and

$$\Im(\zeta_X^2(\sigma + iT)) = -\sum_{n=1}^{X} \frac{d(n)}{n^\sigma} \sin(T \log n).$$

By a generalization of Descartes' Rule of Signs (see Pólya and Szegö [14], Volume II, Part V, Chapter 1, No. 77), the number of real zeros of $\Im(\zeta_X^2(\sigma + iT))$ in the interval $[\alpha, \beta]$ is less than or equal to the number of nonzero coefficients $d(n) \sin(T \log n)$. Hence, the number of zeros of $\Im(\zeta_X^2(\sigma + iT))$ is at most $X - 1$.

Since by assumption the line $t = T$ does not coincide with any zero of $\zeta_X^2(s)$, the zeros of $\Im(\zeta_X^2(s))$ situated along the top edge of R must be real. Furthermore, since the change in argument between two consecutive zeros is at most π, it follows that

$$|\triangle_{[\alpha,\beta]} \arg \zeta_X^2(\sigma + iT)| \le \pi(X - 1).$$

Hence,

$$\triangle_{[\alpha,\beta]} \arg \zeta_X^2(\sigma + iT) = O(X). \tag{16}$$

Finally, along the left edge of R,

$$\zeta_X^2(\alpha + it) = \left(1 + \frac{1 + \displaystyle\sum_{n=2}^{X-1} \frac{d(n)}{n^{\alpha+it}}}{\dfrac{d(X)}{X^{\alpha+it}}} \right) \frac{d(X)}{X^{\alpha+it}},$$

from which it follows that

$$\begin{aligned} \triangle_{[0,T]} \arg \zeta_X^2(\alpha + it) = \triangle_{[0,T]} \arg \left(1 + \frac{1 + \displaystyle\sum_{n=2}^{X-1} \frac{d(n)}{n^{\alpha+it}}}{\dfrac{d(X)}{X^{\alpha+it}}} \right) \\ + \triangle_{[0,T]} \arg \frac{d(X)}{X^{\alpha+it}}. \end{aligned} \tag{17}$$

In virtue of

$$\begin{aligned}\triangle_{[0,T]} \arg \frac{d(X)}{X^{\alpha+it}} &= \triangle_{[0,T]} \arg \frac{d(X)}{X^{\alpha}} e^{-it\log X} \\ &= \triangle_{[0,T]} \arg e^{-it\log X} \qquad (18) \\ &= -T \log X,\end{aligned}$$

it remains for us to examine the first term on the right-hand side of (17), and this is possible on the basis of

$$\frac{d(X)}{X^{\alpha}} > \sum_{1\le n\le X-1} \frac{d(n)}{n^{\alpha}}$$

(in the proof of Theorem 4) that, for any real number t,

$$\left| \frac{1 + \sum_{n=2}^{X-1} \frac{d(n)}{n^{\alpha+it}}}{\frac{d(X)}{X^{\alpha+it}}} \right| < 1.$$

Hence,

$$\triangle_{[0,T]} \arg \left(1 + \frac{1 + \sum_{n=2}^{X-1} \frac{d(n)}{n^{\alpha+it}}}{\frac{d(X)}{X^{\alpha+it}}} \right) = O(1). \qquad (19)$$

Inserting estimates (18) and (19) into (17), we obtain

$$\triangle_{[0,T]} \arg \zeta_X^2(\sigma + it) = -T \log X + O(1). \qquad (20)$$

Finally, inserting estimates (14)–(16), and (20) into (13), we obtain precisely the statement of Theorem 5. Hence, the proof is finished.

5 Final Comments and Suggestions for Future Research

There are gaps in our investigation of the zeros of $\zeta_X^2(s)$ thus far that need to be filled and other directions that must be explored. Below, we identify several research problems that arise naturally from our current investigation.

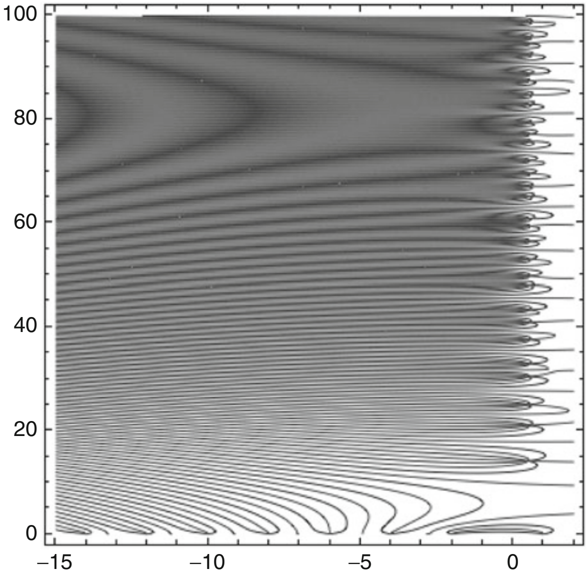


Fig. 1 Contour plot of $\zeta^2(s)$ with $-15 \le \sigma \le 2$ and $0 \le t \le 100$

Problem 1. The proof of Theorem 4 is constructed around the fact that for σ far enough to the left of the origin, one term of $\zeta_X^2(s)$ dominates the rest, as shown in (10). While this is a safe way to ensure the non-vanishing of $\zeta_X^2(s)$, it is far from being most efficient. It would be useful to obtain computational or heuristic evidence that may lead to conjectures for further research that could give an indication of how far from the best possible zero-free regions the bounds in Theorems 3 and 4 really are.

Problem 2. The real and imaginary parts of $\zeta(s)$ are each real-valued functions. The zeros of $\zeta(s)$ are points in the plane where both $\Re(\zeta(s)) = 0$ and $\Im(\zeta(s)) = 0$. These are exactly points where the two level curves cross. This also applies to the real and imaginary parts of $\zeta_X^2(s)$.

Since the zeros of both $\zeta(s)$ and $\zeta_X^2(s)$ are symmetric with respect to the real axis, the contour plots in Figs. 1 and 2 display only the behavior near the critical line $\sigma = 1/2$ and the precise definition of the zeros of $\zeta^2(s)$ and $\zeta_X^2(s)$ in the upper half-plane, respectively. The picture in Fig. 1 shows that the zeros of $\zeta^2(s)$ are clustered near $\sigma = 1/2$.

On the other hand, Fig. 2 shows a striking picture of the zeros of $\zeta_X^2(s)$. Most of the zeros of $\zeta_X^2(s)$ are clustered near $\sigma = 1/2$. What is remarkable is the short strip of zeros near $\sigma = 1/2$. One may inquire about the presence of the trails of zeros of $\zeta_X^2(s)$ in Fig. 2. How far to the left of $\sigma = 1/2$ do these trails of zeros extend? Furthermore, do they become periodic further up in the upper half-plane? The picture in Fig. 2 is compelling. It suggests a phenomenon in need of explanation.

For completeness, it is also worthwhile to conduct a similar investigation of the zeros of the tail of $\zeta_X^2(s)$.

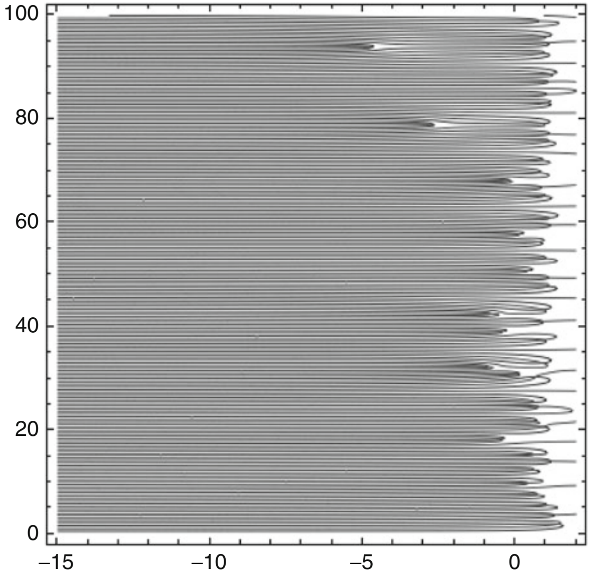


Fig. 2 Contour plot of $\zeta_{20}^2(s)$ with $-15 \le \sigma \le 2$ and $0 \le t \le 100$

Problem 3. The clustering of the zeros of $\zeta(s)$ near $\sigma = 1/2$, first proved by Bohr and Landau [2], is the strongest existing evidence of the Riemann Hypothesis. (See also Titchmarsh [18], Chapter XI, pp. 292–311.) In 1975, Levinson [10] showed that, for any complex a, the zeros of $\zeta(s) - a$ cluster at $\sigma = 1/2$. Specifically, Levinson proved that the number of zeros in the rectangle $|\sigma - 1/2| \le \delta$ and $1 \le t \le T$ of the equation $\zeta(s) = a$ is $(T/2\pi)\log T + O_\delta(T)$, whereas the number of zeros for which $|\sigma - 1/2| > \delta$ and $1 \le t \le T$ is $O_\delta(T)$ at most. The second estimate can be strengthened if $a = 0$. Hence, the clustering of zeros to $\sigma = 1/2$ is more substantial than the clustering of a-values for $a \ne 0$.

Proceeding in exactly the same way as in [6, 10] and using the machinery developed therein, can one determine the locations where the zeros of $\zeta_X^2(s) - a$ are clustered at? Furthermore, is there anything special about the case $a = 0$?

Problem 4. The methods and ideas from [6] and the present paper extend naturally to the smoothed partial sums

$$\sum_{n=1}^{X} \left(1 - \frac{n}{X}\right)^k \frac{d(n)}{n^s}$$

for $k \ge 1$, which provides a better approximation to $\zeta^2(s)$ in the critical strip. What can be said about the distribution of zeros these partial sums?

Acknowledgements The authors wish to express their sincere gratitude to the anonymous referee for carefully reading the original version of this paper and for providing several very helpful comments and suggestions, including a key upper bound for the number of divisors function that greatly simplifies parts of the proofs of Theorems 3 and 4. The research by Kathryn Crosby, Jordan Eliseo, and David Mazowiecki was conducted as part of the 2013 Research Experiences for Undergraduates at The University of Tennessee at Chattanooga that was supported by the National Science Foundation Grant DMS-1261308.

References

1. Alfhors LV (1979) Complex analysis. An introduction to the theory of analytic functions of one complex variable, 3rd edn. McGraw-Hill, New York
2. Bohr H, Landau E (1914) Ein Satz über Dirichlet Reihen mit Andwendung auf die ζ-Funktion und die L-Funktionen. Rend di Palermo 37:269–272
3. Borwein P, Fee G, Ferguson R, van der Waall A (2007) Zeros of partial sums of the Riemann zeta function. Exp Math 16(1):21–40
4. Davenport H (2000) Multiplicative number theory. In: Montgomery HL (ed) Graduate studies in mathematics, vol 74, 3rd edn. Springer, New York
5. Gonek SM (2012) Finite Euler products and the Riemann hypothesis. Trans Am Math Soc 364(4):2157–2191
6. Gonek SM, Ledoan AH (2010) Zeros of partial sums of the Riemann zeta-function. Int Math Res Not 10:1775–1791
7. Hardy GH, Littlewood JE (1923) The approximate functional equation in the theory of the zeta-function, with applications to the divisor-problems of Dirichlet and Piltz. Proc Lond Math Soc 21(2):39–74 (reprinted as pp 174–209 in Collected Papers of GH Hardy, vol II (edited by a committee appointed the London Mathematical Society). Clarendon Press, Oxford, 1967)
8. Hardy GH, Littlewood JE (1929) The approximate functional equation for $\zeta(s)$ and $\zeta^2(s)$. Proc Lond Math Soc 29(2):81–97 (reprinted as pp 213–229 in Collected Papers of GH Hardy, vol II (edited by a committee appointed the London Mathematical Society). Clarendon Press, Oxford, 1967)
9. Ingham AE (1932) The distribution of prime numbers. Cambridge tracts in mathematics and mathematical physics, vol 30 (with a foreword by RC Vaughan). Cambridge University Press, London
10. Levinson N (1975) Almost all roots of $\zeta(s) = a$ are arbitrarily close to $\sigma = 1/2$. Proc Natl Acad Sci USA 72:1322–1324
11. Montgomery HL (1983) Zeros of approximations to the zeta function. Studies in pure mathematics. Birkhäuser, Basel, pp 497–506
12. Montgomery HL, Vaughan RC (2001) Mean values of multiplicative functions. Period Math Hungar 43:199–214
13. Nicolas LJ, Robin G (1983) Explicit estimations for the number of divisors of N. Canad Math Bull 26(4):485–492
14. Pólya G, Szegö G (1998) Problems and theorems in analysis. Theory of functions, zeros, polynomials, determinants, number theory, geometry, vol II (translated from the German by CE Billigheimer; reprint of the 1976 English translation). Classics in mathematics. Springer, Berlin
15. Sándor J, Mitrinović DS, Crstici B (2006) Handbook of number theory, I, 2nd edn. Springer, Dordrecht
16. Spira R (1966) Zeros of sections of the zeta function, I. Math Comput 20:542–550
17. Spira R (1968) Zeros of sections of the zeta function, II. Math Comput 22:168–173
18. Titchmarsh EC (1986) The theory of the Riemann zeta-function, 2nd edn (revised by DR Heath-Brown). The Clarendon Press/Oxford University Press, New York
19. Turán P (1948) On some approximative Dirichlet-polynomials in the theory of the zeta-function of Riemann. Danske Vid Selsk Mat Fys Medd 24(17):36

20. Turán P (1959) Nachtrag zu meiner Abhandlung: "On some approximative Dirichlet-polynomials in the theory of the zeta-function of Riemann". Acta Math Acad Sci Hungar 10:277–298
21. Turán P (1960) A theorem on Diophantine approximation with application to the Riemann zeta-function. Acta Sci Math (Szeged) 21:311–318
22. Turán P (1963) Untersuchungen über Dirichlet-Polynome, Bericht von der Dirichlet-Tagung. Akademie, Berlin, pp 71–80
23. Usol'cev LP (1985) On the estimation of a multiplicative function (Russian). Interuniv. Collect. Sci. Works, Kujbyshev, pp 34–37
24. Voronin SM (1974) The zeros of partial sums of the Dirichlet series for the Riemann zeta-function. Dokl Akad Nauk SSSR 216:964–967 (trans. Soviet Math. Doklady 15:900–903, 1974)

Analysis of Resonance Frequencies for the Problem of Induced Vibrations Along the Human Arm

Irina Viktorova, Lauren Holden, and Sara Bailey Stocks

1 Background

The human body has a very intricate and complicated design. Because of this, it is extremely challenging to model the movements, interactions, and reactions of the human body, especially parts as complicated as the human arm. One challenge that is faced is the fact that the bones of the human body have complicated shapes. This leads to difficulties when it comes to formulating a model that fits the shapes and corresponding interactions and reactions. Another serious problem is that the exact specifications of human biomaterial (tissue) have not been fully discovered. There is plenty of experimental data on bone structure, but it is hard to compare and validate the results of these studies because of the differences in the testing conditions, animal species, and other constraints.

One specific scenario that can be modeled mathematically is the interaction between the human arm and mechanical vibrations. This problem is the topic of [4], applying the method of finite element analysis to human hand arm vibrations. According to [4], when high-energy mechanical vibrations propagate through the human arm, pathological problems are often introduced. This issue is very commonly seen and experienced by people who have an occupation in which they are routinely exposed to machine vibrations. These vibrations and problems most often come from tools such as jackhammers, hand-saws, and drills. Surprisingly, these tools are used in a wide variety of occupations, ranging from construction work to the procedures done by dentists and surgeons. Because of the wide range of people affected by this problem, it is very important to be able to model this interaction in order to better understand it. Long-term exposure to these machines

I. Viktorova (✉) • L. Holden • S.B. Stocks
Department of Mathematical Sciences, Clemson University, Clemson, SC, USA
e-mail: iviktor@clemson.edu

J. Rychtář et al. (eds.), *Collaborative Mathematics and Statistics Research*, Springer Proceedings in Mathematics & Statistics 109, DOI 10.1007/978-3-319-11125-4_7

and their vibrations can lead to Hand-Arm Vibration Syndrome (HAVS). This is a serious condition that is diagnosed in as many as one in every ten people who work with these tools on a regular basis. HAVS is classified by changes in sense of touch of the affected nerves, which oftentimes leads to muscle pain and weakness, irreversible numbness of fingers, and bouts of white finger (also known as Raynaud's phenomenon). If one notices the problem early enough and stops using these tools, he/she may be able to recover from the symptoms; however, most people are not able to catch the problem in time, and they are left with permanent damage [3]. Because of the severity of this syndrome, it is extremely important to be able to model the effect of these resonance frequencies on the human arm so that the problem can be better understood and possibly even prevented.

2 Methodology

The human arm is modeled as a system of two homogeneous viscoelastic rods, joined by the hinge. Viscoelastic material is a compound that has both elastic and viscous properties. Modeling the human arm as a viscoelastic rod means it has both elastic and viscous properties, and it also demonstrates a time-dependent strain. The modeling of the muscle and bone tissue as the homogeneous and viscoelastic solid is based on the experimental mechanical testing data of the bone structure. This model shows that the bone tissue behaves as the linear viscoelastic material. The pressure that surrounds the bone muscle tissue can cause an increase in viscoelasticity and is accounted for by the parameters in the model equation.

The rods are considered homogeneous since the wavelengths (tens of meters), at the frequencies considered, are overwhelmingly greater than the maximum cross-sectional measurements for the human arm. This results in the propagation of the plane (2D—front only), and thus non-homogeneity can be neglected. The viscoelastic parameters can be obtained by running special experimental programs on the bone specimens. The specimens are submerged into viscoelastic media, which models the biologically soft tissue.

The problem of vibrations propagating along the human arm is divided into two parts. The first part is related to the longitudinal oscillations propagating from the hand to the elbow, where η_e is the frequency of vibrations at the elbow. The second problem is related to the propagation of the superposition of longitudinal and bending vibrations (from the elbow to the shoulder) induced at the hand with frequency η_0, with the fraction ratio being defined by the angle of bending for the elbow joint. Therefore, if the angle is zero (straight arm), there is no bending mode involved. Alternatively, if the angle is 90°, then it is only important to account for the bending component.

For some arbitrary value beta between 0° and 90° ($0° < \beta < 90°$), the longitudinal vibrations will be defined by the force components ($\mathrm{P(t)} \cos \beta$) and the bending components (P(t) sin β). P(t) is defined as the value of the load at the elbow, which is obtained from the solution of the longitudinal wave propagation from the hand to the elbow. The displacement equations in this model are derived with no bending mode involved, that is with $\beta = 0°$.

The combined propagation of longitudinal and bending vibrations along the viscoelastic rod of finite length can be modeled by the combined theory of wave propagation in viscoelastic solids [1, 3, 4].

The results for the displacements are given by the following formula, where t is time, x is the location of the cross-section, λ is the length of the rod, ω_1 is the vibrational frequency, ψ is the phase angle, and α is the viscoelastic parameter.

$$\frac{u(x,t)}{u_0} = \left[\frac{\cosh(2\frac{\lambda-x}{C}A) + \cos(\frac{\lambda-x}{C}B)}{\cosh(2\frac{\lambda}{C}A) + \cos(\frac{\lambda}{C}B)}\right]^{\frac{1}{2}} * \cos(\omega_1 t - \psi) \qquad (1)$$

where

$$A = \sqrt{\frac{\omega_1}{2}}\left[\left[x^2{\omega_1}^{2\alpha} - 2x{\omega_1}^{1+\alpha}\cos(\frac{\pi}{2}(1+\alpha)) + {\omega_1}^2\right]^{\frac{1}{2}}\right.$$
$$\left. +x{\omega_1}^2\cos(\frac{\pi}{2}(1+\alpha)) - \omega_1\right]^{\frac{1}{2}}$$
$$B = \sqrt{\frac{\omega_1}{2}}\left[\left[x^2{\omega_1}^{2\alpha} - 2x{\omega_1}^{1+\alpha}\cos(\frac{\pi}{2}(1+\alpha)) + {\omega_1}^2\right]^{\frac{1}{2}}\right.$$
$$\left. -x{\omega_1}^2\cos(\frac{\pi}{2}(1+\alpha)) - \omega_1\right]^{\frac{1}{2}}$$
$$\psi = \arctan\left[\frac{\sin(\frac{x}{C}B)\sinh(\frac{2\lambda-x}{C}A) + \sin(\frac{2\lambda-x}{C}B)\sinh(\frac{x}{C}A)}{\cos(\frac{x}{C}B)\cosh(\frac{2\lambda-x}{C}A) + \cos(\frac{2\lambda-x}{C}B)\cosh(\frac{x}{C}A)}\right]$$

The amplitude of the bending vibrations is given by the following formula

$$W = \frac{\omega_0}{2}\left[\frac{S_1\sin(kx) - S_2\sinh(kx)}{\sin(k\lambda)\cosh(k\lambda) - \cos(k\lambda)\sinh(k\lambda)} + \cos(kx) + \cosh(kx)\right] \qquad (2)$$

where

$$S_1 = 1 - \cos(k\lambda)\cosh(k\lambda) - \sin(k\lambda)\sinh(k\lambda)$$
$$S_2 = 1 - \cos(k\lambda)\cosh(k\lambda) + \sin(k\lambda)\sinh(k\lambda)$$

and $k = \frac{r}{2}$ for the cylinder of radius r.

3 Results

Using these relations, we can calculate the dependence of the vibration's amplitude on time "t" for the arbitrary cross-section x. Setting $x = \lambda_1$, we obtain:

$$\frac{u(\lambda_1, t)}{u_0} = \left[\frac{2}{\cosh(2\frac{\lambda_1}{C}A) + \cos(2\frac{\lambda_1}{C}B)}\right]^{\frac{1}{2}} * \cos(\omega_1 t - \psi). \tag{3}$$

The above equation defines the displacement or amplitude of longitudinal vibrations at the elbow, with the parameters of the induced vibrations located at the hand. The same equation also defines one of the boundary conditions for the problem of bending vibrations propagating from the elbow to the shoulder. The second boundary condition for the part of the arm from the elbow to the shoulder joint can be formulated based on the position of the arm. The analysis of the problem can be simplified with the assumption of no constraint at the end of the rod $x = \lambda_2$. More accurate boundary conditions require more detailed experimental data analysis.

It should be emphasized that the superposition of longitudinal and bending vibrations along the second rod that models the part of the arm above the elbow and below the shoulder is characterized by the amplitude, which depends on the angle of the elbow. The amplitude of the bending vibrations at $x = \lambda_2$ is defined by the formula:

$$W|(x = \lambda) = \omega_0 * \left[\frac{sin(k\lambda) - \sinh(k\lambda)}{\sin(k\lambda)\cosh(k\lambda) - \cos(k\lambda)\sinh(k\lambda)}\right]. \tag{4}$$

Furthermore, the relationship between the amplification factor ρ for bending vibrations and the corresponding frequencies ω is given by the equation:

$$\rho = \frac{(cosh(\omega) - \cos(\omega) - \sin(\omega)\sinh(\omega))^{\frac{1}{2}}}{\sqrt{2}(\sinh(\omega) - \sin(\omega))}. \tag{5}$$

According to the above equations, the ratio of the amplitudes is proportional to the squares of the ratios of the frequencies. This fact will be used later in the analysis of the resonance frequencies.

Analysis of Eq. (5) allows one to model and predict the resonance frequencies of the described phenomenon. The first step is in taking the first derivative of $\rho(\omega)$ and setting it equal to zero. This gives an equation that can be solved for values of ω, which correspond to the frequencies that yield the maximum/minimum amplitudes. The first derivative is given by:

$$\frac{d\rho}{d\omega} = \frac{[\sinh(\omega) - \sin(\omega)] * [\sinh(\omega) + \sin(\omega) - \cos(\omega)\sinh(\omega) - \sin(\omega)\cosh(\omega)]}{2\sqrt{\cosh(\omega) - \cos(\omega) - \sin(\omega)\sinh(\omega)}}$$

$$-\frac{[\cosh(\omega)-\cos(\omega)]*\sqrt{\cosh(\omega)-\cos(\omega)-\sin(\omega)\sinh(\omega)}}{[\sinh(\omega)-\sin(\omega)]^2}. \tag{6}$$

Setting Eq. (6) equal to zero and solving verifies that the first extremum is $\omega = 3\pi$. Using this value of ω to simplify Eq. (6), we get:

$$\frac{d\rho}{d\omega} = \sinh^2(\omega) * [\sin(\omega) - \cos(\omega) - 1] = 0. \tag{7}$$

The solution demonstrates that the maximum/minimum amplitude occurs at frequencies given by:

$$\omega = \frac{\pi}{2} + n\pi, n \in \mathbb{N}$$

and

$$\omega = \pi + n\pi, n \in \mathbb{N} \tag{8}$$

The second derivative test determines whether the amplitude corresponding to the first extremum frequency of $\omega = 3\pi$ is a maximum or minimum. The second derivative is given by:

$$\frac{d^2\rho}{d\omega^2} = \sinh(\omega) * [2\sin(\omega) - 2\cos(\omega) - 2 + \sinh(\omega)[\cos(\omega) + \sin(\omega)]]. \tag{9}$$

Plugging in $\omega = 3\pi$ gives a value of −38,388,000, which is negative. Thus, this frequency corresponds to a maximum amplitude. Plugging $\omega = 3\pi$ into Eq. (5) gives a value of $\rho = 0.008984$ for the maximum amplitude. Furthermore, as the value of $\sinh(\omega)$ gets large the sign (positive or negative) of the second derivative depends on the expression $[\cos(\omega) + \sin(\omega)]$. Thus, the maximum amplitudes occur at $\omega = 3\pi, 5\pi, 7\pi, 9\pi, \ldots$ because $[\cos(\omega) + \sin(\omega)]$ is negative at these values. Then, the fact that the ratio of the amplitudes is proportional to the ratio of the squares of the frequencies, as follows from Eq. (5), allows to conclude that the first ratio of the maximum amplitudes is $\frac{9\pi^2}{25\pi^2} = \frac{9}{25}$.

Verification of the obtained numerical results is implemented by comparative analysis with the experimental data, under the assumption that parameter α is very close to 1. (The analysis of the experimental data with the bone tissue samples tested at various loading rates confirmed by the value of $\alpha = 0.98$.) Figure 1 demonstrates the dependence of amplitude on frequency for the propagation of bending vibrations at the elbow of the human operator, where η_e denotes the frequency of vibrations at the elbow, and η_0 denotes the applied frequency of the inducer. The graph shows that the ratio of the frequencies for the two distinct maximums is very close to $\frac{9}{25}$, which verifies the results of the model analysis. The local minimum occurring at approximately 5π requires further investigation. It should be noted that Fig. 1 was constructed using data from an experimental

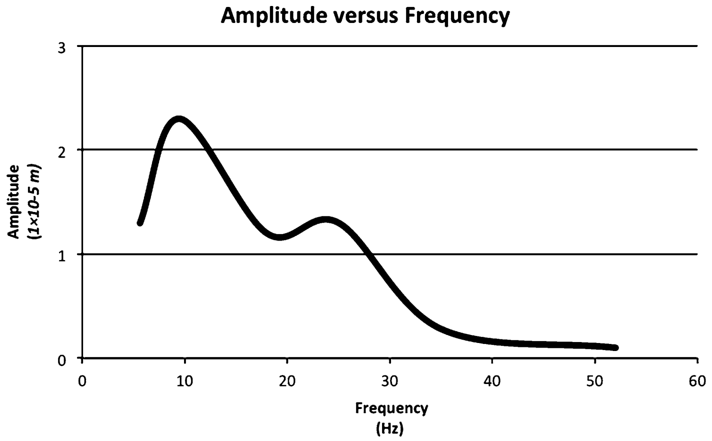


Fig. 1 This figure shows the dependence of bending amplitude on the frequency of induced mechanical vibrations along the human arm

study [2] that was carried out independently from the model approach discussed in this paper, which is significant for the verification of the mathematically modeled results for the resonance frequencies.

Conclusion

The analysis discussed in this paper led to a model of the effect of resonance frequencies on the human arm. The results of this analysis are verified by experimental data, which was collected independently. From the experimental data, Fig. 1 was created, which verified the results of the model approach.

It should be emphasized that the model approach examined in this study allows for the determination of the viscoelastic material parameters, as well as the value of the stress propagation rate. These parameters cannot be obtained from the samples of the human bone tissue due to the specifics of manufacturing and size limitations.

Experimental data is difficult to obtain from human subjects because of the possible harm to humans undergoing the experiment. The results obtained on the postmortem bone and soft tissue specimens do not present accurate parameters and are often useless. Therefore, the theoretical analysis for the propagation of the mechanical vibrations along the human arm establishes in particular the values of the resonance frequencies. This is of upmost importance because it provides valuable information on safety and overall advanced operations development. As stated earlier, this problem affects

(continued)

people in a wide range of occupations, and it is a very common and serious problem. Long-term exposure to this problem can lead to Hand-Arm Vibration Syndrome, which oftentimes leaves people with irreversible damage. Thus, this model and the corresponding analysis are extremely important and practical.

References

1. Christensen RM (1971) Theory of viscoelasticity: an introduction. Academic, New York. Print
2. Frolov KV, Potemkin BA (1972) Nonlinear vibrations and transitional processes in machines: experimental study of human response to the induced mechanical vibrations. Nauka, Moscow. Print
3. Kenny T (2014) Hand-arm Vibration Syndrome. Patient.co.uk. N.p., 02 Aug. 2012. Web. 22 Mar
4. Zadpoor AA (2006) Finite element method analysis of human hand arm vibrations. Int J Sci Res 16:391–395

Modeling Risky Sexual Behavior Among College Students: Predictors of STD

Qi Zhang, Haseeb Kazi, and Sat Gupta

1 Introduction

A risky sexual behavior is one that increases one's risk of contracting sexually transmitted infections and experiencing unintended pregnancies. Unprotected sexual activities, including high level of non-condom use could increase the risk of HIV/STDs transmission [3, 9]. Risky sexual behavior usually includes having more than one sexual partner, changing sexual partners frequently, having oral, vaginal, or anal sexual contact without a condom, and using unreliable methods of birth control, or using birth control inconsistently [1, 2]. As the youth of today start engaging in sexual practice at earlier ages, the incidence of sexually transmitted diseases has been rising in recent years. A study has shown that average age of first sexual intercourse now is 16.7 years, and first sexual intercourse is significantly associated with human papillomavirus (HPV) infection [6]. The Youth Risk Behavior Surveillance System (YRBSS) reported in 2011 that sexual behavior that contributes to unintended pregnancy and STDs is one of six categories of priority health-risk behaviors among youth and young adults [2]. Previous sexually transmitted disease surveys among young adults show that they account for 50 % of

Q. Zhang
Department of Chemistry and Biochemistry, The University of North Carolina
at Greensboro, Greensboro, NC 27402, USA
e-mail: q_zhang@uncg.edu

H. Kazi
Department of Medicine, Emory University, Atlanta, GA 30322, USA
e-mail: hkazi@emory.edu

S. Gupta (✉)
Department of Mathematics and Statistics, The University of North Carolina
at Greensboro, Greensboro, NC 27402, USA
e-mail: sngupta@uncg.edu

J. Rychtář et al. (eds.), *Collaborative Mathematics and Statistics Research*, Springer
Proceedings in Mathematics & Statistics 109, DOI 10.1007/978-3-319-11125-4_8

the 19 million new STD cases each year, but they just account for only 25 % of the sexually active population [8]. In these young adults, college students account for a large proportion. STDs can affect students' health and their future. It is important to help students understand the basics of STDs. However, studies show that even though college students have relevant knowledge about STDs, they still continue to engage in risky sexual behaviors [7]. That means students' sexual education needs to be strengthened, including the harmful implications of STDs, as well as appropriate intervention measures. Although many researchers have studied STDs in adolescents, the studies of the predictors of STD incidences are less common. Analyzing factors that affect sexual behavior among American college students could provide the scientific basis for formulation of sexual health education.

Since the sexually transmitted disease questions are sensitive to many students, students may not provide truthful responses. In the previous study by Spears-Gill et al. [4], a survey was conducted about students' sexual behaviors at The University of North Carolina at Greensboro (UNCG). It was observed that students tend to significantly underreport risky sexual behaviors. In the present study, we try to quantify the extent of under-reporting, and also study the relationship between STD history and some important predictors, such as age, gender, and number of sexual partners. This study confirms our expectation that those students who have more sexual partners are more likely to be diagnosed with STDs. It also confirms another expectation that there will be higher incidence of STD in women than men because of their physiological structure. The skin of vagina is thinner and more delicate than the skin of penis, that means bacteria and viruses are easier to penetrate. And also the vagina provides a moist environment for bacteria to grow. Previous studies have shown that 13 % of females and 4 % of males have had STD [5].

2 Sample Characteristics

The target population in this study is undergraduate students enrolled at UNCG during the 2012–2013 academic year. Students offered to take part in the study voluntarily and filled out the survey questionnaires during regular mathematics and statistics class times. Subjects were 31.4 % male and 68.6 % female (392 valid reporting), 3.7 % were married and 95.8 % were not married (409 valid reporting). They had a mean age of 19.99 (411 valid reporting), and 90.3 % subjects were in the 18–22 age group. The distribution of their class levels was: 47.6 % freshman, 32.0 % sophomore, 12.7 % junior, 7.3 % senior, and 0.5 % other (410 valid reporting). Proper IRB approval was sought and received before data collection.

3 Procedures

The researchers asked the subjects questions using two survey methods. One method was face-to-face interview, which meant the subject answered questions directly to the researcher; and the other method was a check-box method which meant the subject completed the survey completely anonymously. Before the survey, students received basic information about the study, such as risks, benefits, and the questions that would be on the questionnaire. Then the subjects completed a consent form and a demographic information form. The two sensitive questions about sexual behavior used in the survey were "**How many sexual partners have you had in the last 12 months?**" and "**Have you ever been told by a healthcare professional that you have a sexually transmitted disease?**" The face-to-face interview group answered sexual behavior questions directly to the interviewer, and the check-box group completed the questionnaires anonymously and dropped them in a box. Professors under whose guidance the study was completed were available during data collection.

4 Results

The collected data was analyzed by using SPSS software. In the first part of the analysis, the STD incidence (yes or no) was set as the dependent variable, and the survey group (face-to-face and check-box) was set as the independent variable. Logistic regression was used to test under-reporting of STDs, summary is shown in Table 1.

The odds ratio of 0.213 indicates that odds of a yes response with the face-to-face survey method are 79 % smaller as compared with the check-box method, with a p-value of 0.005. This confirms the significant under-reporting of STD in face-to-face interviews.

In the second part of the analysis, we again used STD incidence (yes or no) as the dependent variable, and used age, gender, and number of sexual partners as independent variables. We included the survey group also in the model to control for it. The results in Table 2 below are based on the logistic regression fit using forward selection method. Clearly, the number of sexual partners, age, and gender are significant. Although survey group was not a significant predictor at 5 % level, we kept it in the model to control for it.

Table 1 Under-reporting of STD in Face-to-Face interviews with Logistic Regression Check-Box="1"; Face-to-Face="2"

Predictor	B	SE	Wald	df	Sig.	Exp (B)
Survey group	−1.548	0.557	7.718	1	0.005	0.213
Constant	−0.755	0.689	1.199	1	0.273	0.470

Table 2 Predictors of STD with Logistic Regression Male = "0"; Female = "1"; Check-Box = "1"; Face-to-Face = "2"

Predictor	B	SE	Wald	df	Sig.	Exp (B)
Survey group	−1.102	0.603	3.337	1	0.068	0.332
# of sexual partners	0.430	0.123	12.222	1	0.000	1.537
Gender	2.979	1.308	8.236	1	0.004	19.672
Age	0.180	0.050	13.068	1	0.000	1.197
Constant	−8.402	2.045	16.887	1	0.000	0.000

5 Discussion

The previous study of Spears-Gill et al. [4] showed serious under-reporting of the number of sexual partners and the incidences of STDs during face-to-face interview since this method had no anonymity, at least relative to the researcher. In the present study, the results in Table 1 show that there is 79 % under-reporting of STD incidence when using face-to-face survey as compared to check-box method since the relative risk is 0.213. The results in Table 2 demonstrate that the factors strongly associated with the incidence of STD, after controlling for survey group, are gender, age, and number of sexual partners. Women have 19.67 times the risk men have, with all else kept equal. The risk of STD increases by 53.7 % with each additional sexual partner. The risk increases by 20 % with each additional year of age. However, since most undergraduate students in the study are similar in age, this factor may not really be a significant factor.

Several conclusions emerge from this study. Firstly, there is a serious problem of under-reporting in face-to-face surveys dealing with sensitive questions. Secondly, women are more susceptible to STDs than men are. The physiological structures are easier to allow women to get infections under the same conditions. Thirdly, the more sexual partners one has, greater the risk of having STDs.

As mentioned before, healthy sexual behaviors for students' physical and psychological development are very important. In order to reduce the incidence of sexually transmitted disease among college students, high schools should strengthen students' sexual education program, including basic knowledge of STDs and some prevention. Even though this study successfully demonstrated students' under-reporting of sexual behaviors and the predictors of STD, it had several limitations. The survey had some missing data, including gender, number of sexual partners, and the incidences of STD. Those missing data may affect our conclusions. Also, the survey is conducted at one college campus located in the so-called "Bible Belt". The results may not reflect sexual behaviors of the entire American college student population. This survey can only provide some directions for future studies on this topic.

References

1. Eaton DK, Kann L, Kinchen S, Shanklin S, Ross J, Hawkins J, Harris WA, Lowry R, McManus T, Chyen D et al (2009) Youth risk behavior surveillance—United States, 2009. Morb Mortal Weekly Rep 59(5):1–142
2. Eaton DK, Kann L, Kinchen S, Shanklin S, Ross J, Hawkins J, Harris WA, Lowry R, McManus T, Chyen D et al (2011) Youth risk behavior surveillance—United States, 2011. Morb Mortal Weekly Rep 61(4):1–162
3. El-Bassel N, Wingood G, Wyatt GE, Jemmott JB 3rd, Pequegnat W, Landis JR, Bellamy S, Gilbert L, Remien RH, Witte S et al (2010) Risky sexual behavior and correlates of std prevalence among african american hiv serodiscordant couples. AIDS Behav 14(5):1023–1031
4. Gill TS, Tuck A, Gupta S, Crowe M, Figueroa J (2013) A field test of optional unrelated question randomized response models: estimates of risky sexual behaviors. In: Topics from the 8th annual UNCG regional mathematics and statistics conference. Springer, Berlin, pp 135–146
5. Hahm HC, Lee J, Ozonoff A, Amodeo M (2007) Predictors of stds among asian and pacific islander young adults. Perspect Sexual Reproduct Health 39(4):231–239
6. Kahn JA, Rosenthal SL, Succop PA, Ho GYF, Burk RD (2002) Mediators of the association between age of first sexual intercourse and subsequent human papillomavirus infection. Pediatrics 109(1):e5
7. Lewis JE, Malow RM, Ireland SJ (1997) HIV/AIDS risk in heterosexual college students: a review of a decade of literature. J Am College Health 45(4):147–158
8. Weinstock H, Berman S, Cates W (2004) Sexually transmitted diseases among American youth: incidence and prevalence estimates, 2000. Perspect Sexual Reproduct Health 36(1):6–10
9. Xia G, Yang X (2005) Risky sexual behavior among female entertainment workers in China: implications for HIV/STD prevention intervention. AIDS Educat Prevent 17(2):143–156

Modeling Asian Carp Invasion Using Evolutionary Game Theory

Jasmine Everett, Marwah Jasim, Hyunju Oh, Jan Rychtář, and Hakimah Smith

1 Introduction

Asian Carp were imported from China in 1970s to improve water quality of aquaculture ponds in the Mississippi river along Illinois. The fish can grow incredibly quickly and can weigh up to 150 pounds and reach an average size of about 30–40 inches. They could eat 5–20 % of their body weight each day, so the species seemed to be an ideal option to control aquatic vegetation. However, having no natural predators, the population of Asian Carp grew exponentially, threatening the native fish whose diet overlaps with the Asian Carp's diet. The population of native fish now decreases quickly in the upper Mississippi river system [10]. Consequently, Asian Carp are now considered invasive species, highly detrimental to the ecological balance. In April 2012 Congress enacted the "Stop Invasive Species Act" [3]. The act requires the U.S. Corp of Engineers to implement measures to prevent Asian Carp from invading the Great Lakes from the Mississippi through the Chicago area canal system. Also, the Obama Administration released the 2013 Asian Carp Control Strategy Framework [5].

The objective of this paper is to model the interaction between native species and Asian Carp. In Sect. 2 we develop a game-theoretical model evaluating the costs and benefits of the interactions between native predator and prey fish and the invasive Asian Carp species. In Sect. 3 we provide the analysis of the model. More

J. Everett • M. Jasim • H. Oh (✉) • H. Smith
Bennett College, 900 E. Washington St, Greensboro, NC 27401, USA
e-mail: jasmine.everett@bennett.edu; marwah.jasim@bennett.edu; hoh@bennett.edu; hakimah.smith@bennett.edu

J. Rychtář
The University of North Carolina at Greensboro, Greensboro, NC 27412, USA
e-mail: rychtar@uncg.edu

J. Rychtář et al. (eds.), *Collaborative Mathematics and Statistics Research*, Springer Proceedings in Mathematics & Statistics 109, DOI 10.1007/978-3-319-11125-4_9

specifically, in Sect. 3.1 we study the conditions under which the native species can coexist (without the presence of Asian Carp). In Sects. 3.2 and 3.3 we study the conditions under which the Asian Carp cannot invade the native fish population. Finally, in Sect. 4 we summarize the findings from previous sections and provide recommendations for potential control measures.

2 Mathematical Model

To build a model of interactions between Asian Carp and the native species of fish, we will focus on three specific species: *Silver Carp* (SC), *Gizzard Shad* (GS), and *Largemouth Bass* (LB). *Silver Carp* is a non-indigenous most invasive species of Asian Carp; *Gizzard Shad* and *Largemouth Bass* are popular native fish in the Upper Mississippi river. Moreover, *Gizzard Shad* is a prey of the predatory *Largemouth Bass* and thus our model can capture the prey–predator interaction common in native fish.

The proportion of each species in the population will be denoted by p_{SC}, p_{GS}, and p_{LB}, respectively, with the sum of the proportions of SC, GS, and LB equals to 1.

We will now define the benefits and costs for SC, GS, and LB individuals. Let R be the value of common resources (such as algae and other microorganism) for all of the three species of fish. All fish consume these shared resources. However, the resources are not consumed equally, but rather the consumption is proportional to the weight of the species [9].

Consequently, a fish of a type $F \in \{SC, GS, LB\}$ will consume $w_F/(p_{SC}w_{SC} + p_{GS}w_{GS} + p_{LB}w_{LB})$ of available resources, where w_F is the average weight of the fish of type F. In the Upper Mississippi river, the average weights in 10^2 pounds are $w_{SC} = 1$, $w_{GS} = 0.05$, and $w_{LB} = 0.2$ [8, 11].

The predator LB also eats GS. This means additional benefits to LB and extra costs to GS. We quantify the benefits to LB by $p_{GS}V_{LB}^{(GS)}$, i.e. as being proportional to the abundance of GS. Here, $V_{LB}^{(GS)}$ is a benefit of catching GS (i.e., a benefit of one GS caught by LB). Similarly, we quantify the costs to GS by $p_{LB}C_{GS}^{(LB)}$, where $C_{GS}^{(LB)}$ is (an average) cost of being caught by LB (more specifically a cost to a population of GS caused by a single LB). Because the cost $C_{GS}^{(LB)}$ for GS represents the fact of being eaten, while the benefits $V_{LB}^{(GS)}$ for LB correspond more to having a snack we may assume that

$$C_{GS}^{(LB)} > 2V_{LB}^{(GS)}. \tag{1}$$

Different species of fish also produce different number of eggs. The average number of eggs (in 10^6 per year) for our three species are $E_{SC} = 4.2$, $E_{GS} = 0.5$, and $E_{LB} = 0.08$ [2, 6, 10]. For a particular fish, the number of produced eggs may depend on the number of consumed resources. However, to better model the fact that *Silver Carp* consumes disproportionately more resources as well as produces much more eggs than the native species, we will consider the benefits of resource consumption and egg production as additive.

Finally, *Silver Carp* population is now controlled by humans using fishing, electronic devices, chemicals, and other mechanisms [7], and it will also move to another river when the density is too high. So we define $C_{SC}^{(H)}$ to be the cost of human control and $p_{SC}C_{SC}^{(M)}$ to be the cost of moving (the SC is more likely to move to another river when their density and thus their proportion is high).

The net benefits (benefits minus costs) to all three species thus are

$$\mathcal{E}_{SC} = \frac{w_{SC}}{p_{SC}w_{SC} + p_{GS}w_{GS} + p_{LB}w_{LB}}R + E_{SC} - C_{SC}^{(H)} - p_{SC}C_{SC}^{(M)}, \tag{2}$$

$$\mathcal{E}_{LB} = \frac{w_{LB}}{p_{SC}w_{SC} + p_{GS}w_{GS} + p_{LB}w_{LB}}R + E_{LB} + p_{GS}V_{LB}^{(GS)}, \tag{3}$$

$$\mathcal{E}_{GS} = \frac{w_{GS}}{p_{SC}w_{SC} + p_{GS}w_{GS} + p_{LB}w_{LB}}R + E_{GS} - p_{LB}C_{GS}^{(LB)}. \tag{4}$$

The model parameters and notation are summarized in Table 1.

Table 1 Notations and parameter values

Notation	Meaning (and value if known)
w_{SC}	An average weight of *Silver Carp* in 10^2 pounds (1)
w_{LB}	An average weight of *Largemouth Bass* in 10^2 pounds (0.2)
w_{GS}	An average weight of *Gizzard Shad* in 10^2 pounds (0.05)
E_{SC}	An average number of eggs produced by SC in millions per year (4.2)
E_{LB}	An average number of eggs produced by LB in millions per year (0.08)
E_{GS}	An average number of eggs produced by GS in millions per year (0.5)
$C_{GS}^{(LB)}$	Cost to GS caused by being caught by LB
$V_{LB}^{(GS)}$	Benefit of caught GS to LB
$C_{SC}^{(H)}$	Cost of human control measures to SC
$C_{SC}^{(M)}$	Cost to SC caused by spreading to another place
p_{SC}	Proportion of SC in the river
p_{LB}	Proportion of LB in the river
p_{GS}	Proportion of GS in the river
$\mathcal{E}_{SC}$	Net benefits (i.e. benefits minus costs) of SC
$\mathcal{E}_{LB}$	Net benefits (i.e. benefits minus costs) of LB
$\mathcal{E}_{GS}$	Net benefits (i.e. benefits minus costs) of GS
R	Resources available in the river

3 Analysis

We are primarily interested in conditions under which GS and LB can coexist without SC and conditions under which SC cannot invade the $GS - LB$ mixture. The payoffs (2)–(4) define a non-linear game and the game will be solved for stable states using standard methods shown, for example in, [1, Chapter 7]. The coexistence condition is $\mathscr{E}_{LB} = \mathscr{E}_{GS}$ (under $p_{SC} = 0$). The non-invadability condition then is $\mathscr{E}_{LB} > \mathscr{E}_{SC}$ (under $p_{SC} = 0$).

3.1 Coexistence of LB and GS

For the stability of the $GS - LB$ mixture, we need

$$\mathscr{E}_{LB} - \mathscr{E}_{GS} = 0, \tag{5}$$

$$\frac{\partial}{\partial p_{LB}}(\mathscr{E}_{LB} - \mathscr{E}_{GS}) < 0, \tag{6}$$

where (5) means that GS does equally well as LB in the mixture and (6) means that even when the population deviates from the equilibrium (by a little bit), the replicator dynamics [1, Chapter 2], [4] will bring it back to the steady state.

When $p_{SC} = 0$, we have that $p_{GS} = 1 - p_{LB}$, and the formulas (3) and (4) become

$$\mathscr{E}_{LB} = \frac{w_{LB}}{w_{GS} + p_{LB}(w_{LB} - w_{GS})} R + E_{LB} + (1 - p_{LB}) V_{LB}^{(GS)}, \tag{7}$$

$$\mathscr{E}_{GS} = \frac{w_{GS}}{w_{GS} + p_{LB}(w_{LB} - w_{GS})} R + E_{GS} - p_{LB} C_{GS}^{(LB)}. \tag{8}$$

Condition (5) is thus equivalent to

$$0 = \frac{w_{LB} - w_{GS}}{w_{GS} + p_{LB}(w_{LB} - w_{GS})} R - (E_{GS} - E_{LB} - V_{LB}^{(GS)}) + p_{LB}(C_{GS}^{(LB)} - V_{LB}^{(GS)}). \tag{9}$$

We now set

$$a = (w_{LB} - w_{GS})(C_{GS}^{(LB)} - V_{LB}^{(GS)}), \tag{10}$$

$$b = w_{GS}(C_{GS}^{(LB)} - V_{LB}^{(GS)}) - (w_{LB} - w_{GS})(E_{GS} - E_{LB} - V_{LB}^{(GS)}), \tag{11}$$

$$c = (w_{LB} - w_{GS})R - w_{GS}(E_{GS} - E_{LB} - V_{LB}^{(GS)}), \tag{12}$$

and so (9) becomes equivalent to

$$0 = ap_{LB}^2 + bp_{LB} + c. \tag{13}$$

We will now focus on the stability condition (6). Because $w_{LB} > w_{GS}$ and, from (1), $C_{GS}^{(LB)} > V_{LB}^{(GS)}$, we get that $a > 0$. This means that the only candidate for a stable mixture of GS and LB is the smaller root of (13), i.e.

$$p_{LB} = \frac{-b - \sqrt{b^2 - 4ac}}{2a}. \tag{14}$$

In order for p_{LB} defined by (14) to be positive, we need $b < 0$ and $c > 0$. It follows from (11) (and the need for $b < 0$) that we need

$$E_{GS} - E_{LB} - V_{LB}^{(GS)} > 0. \tag{15}$$

In fact, if (15) does not hold, then $a, b, c > 0$ and thus $p_{LB} = 1$ is the only stable outcome.

Because we also need $c > 0$, it follows from (12) that we also need

$$(w_{LB} - w_{GS})R > w_{GS}(E_{GS} - E_{LB} - V_{LB}^{(GS)}). \tag{16}$$

We note that when $c < 0$, i.e. when

$$R < \frac{w_{GS}}{w_{LB} - w_{GS}}(E_{GS} - E_{LB} - V_{LB}^{(GS)}) = R_{\min} \tag{17}$$

the only stable outcome of the situation is $p_{LB} = 0$, i.e. over a long period of time, only GS can exist in the population.

To make sure that $p_{LB} < 1$, we look at the vertex $v = -b/2a$ of a quadratic function from (13). If $v > 1$, i.e. if $2a + b < 0$, then all that is further needed is $a + b + c < 0$ (i.e., the quadratic function from (13) is positive at 0 since $c > 0$ and negative at 1 since $a + b + c < 0$). If $v < 1$, i.e. if $2a + b > 0$, then there may be no root and thus the condition needed for a root to exist is $b^2 - 4ac > 0$, i.e. $c < b^2/4a$. Thus, for the stable root of (13) to exist, we need

$$b < 0 < c < \begin{cases} -b - a, & \text{if } 2a + b < 0, \\ b^2/4a, & \text{otherwise.} \end{cases} \tag{18}$$

Consequently, the stable mixture of GS and LB exists only if $R \in (R_{\min}, R_{\max})$ where $R_{\min}$ is given by (17) as it follows from the condition $c > 0$ and $R_{\max}$ is similarly given by conditions on caps on c in (18). Moreover, when $R < R_{\min}$, then $p_{GS} = 1$ (i.e., GS only) is stable and if $R > R_{\max}$, then $p_{LB} = 1$ (i.e., LB only) is stable.

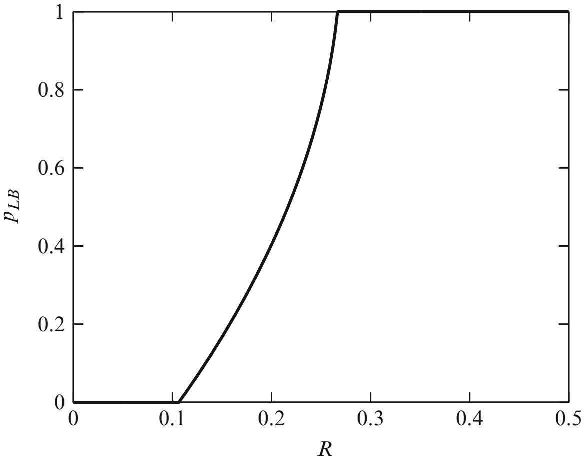


Fig. 1 The stable proportion of LB, p_{LB}, in the population consisting of LB and GS as R varies. The parameter values are $w_{SC} = 1$, $w_{LB} = 0.2$, $w_{GS} = 0.05$, $E_{SC} = 4.2$, $E_{LB} = 0.08$, $E_{GS} = 0.5$, $V_{LB}^{(GS)} = 0.1$, $C_{GS}^{(LB)} = 0.22$, $C_{SC}^{(H)} = 5.5$

We note that the cases on the cap on c from (18) depend on $C_{GS}^{(LB)} - V_{LB}^{(GS)}$. As $C_{GS}^{(LB)} - V_{LB}^{(GS)}$ increases, so does $2a + b$ and thus the cap on c is more likely given by $b^2/4a$. Also, it follows that as $C_{GS}^{(LB)} - V_{LB}^{(GS)}$ increases, $R_{\max}$ decreases.

By (14) and (12),

$$\frac{\partial p_{LB}}{\partial R} = \frac{\partial p_{LB}}{\partial c} \cdot \frac{\partial c}{\partial R} > 0 \tag{19}$$

and thus the proportion of LB in stable GS–LB mixture increases with the increase of available resources R.

See Fig. 1 for the illustration of the results of this section.

3.2 Non-invadability of GS–LB Mixture by SC

Now, assume that GS and LB coexist in a stable mixture (i.e., parameters satisfy (18)). We are interested to find out under what conditions SC cannot invade. That is, we need to find out when

$$\mathscr{E}_{LB} - \mathscr{E}_{SC} > 0 \tag{20}$$

under the assumption that $p_{SC} = 0$ and $p_{LB} \in (0, 1)$ solves (13). Since, by (9),

$$-\frac{R}{w_{GS} + p_{LB}(w_{LB} - w_{GS})} = \frac{-(E_{GS} - E_{LB} - V_{LB}^{(GS)}) + p_{LB}(C_{GS}^{(LB)} - V_{LB}^{(GS)})}{w_{LB} - w_{GS}} \tag{21}$$

we get, using (7) and (2),

$$\begin{aligned}\mathscr{E}_{LB} - \mathscr{E}_{SC} &= -\frac{w_{SC} - w_{LB}}{w_{GS} + p_{LB}(w_{LB} - w_{GS})} R + (1 - p_{LB}) V_{LB}^{(GS)} \\ &\quad - (E_{SC} - E_{LB}) + C_{SC}^{(H)}\end{aligned} \tag{22}$$

$$\begin{aligned}&= \frac{w_{SC} - w_{LB}}{w_{LB} - w_{GS}} \big(p_{LB}(C_{GS}^{(LB)} - V_{LB}^{(GS)}) - (E_{GS} - E_{LB} - V_{LB}^{(GS)})\big) \cdots \\ &\qquad \cdots - p_{LB} V_{LB}^{(GS)} - (E_{SC} - E_{LB}) + C_{SC}^{(H)} + V_{LB}^{(GS)}.\end{aligned} \tag{23}$$

Thus, for the non-invadability condition (20) to hold, we need

$$\begin{aligned}&\frac{w_{SC} - w_{LB}}{w_{LB} - w_{GS}} \big(-p_{LB}(C_{GS}^{(LB)} - V_{LB}^{(GS)}) + (E_{GS} - E_{LB} - V_{LB}^{(GS)})\big) + p_{LB} V_{LB}^{(GS)} \\ &\quad + (E_{SC} - E_{LB}) < C_{SC}^{(H)} + V_{LB}^{(GS)}.\end{aligned} \tag{24}$$

We note that, by (1), and because of the values of w_{SC}, w_{LB}, w_{GS},

$$\frac{\partial}{\partial p_{LB}} (\mathscr{E}_{LB} - \mathscr{E}_{SC}) = \frac{w_{SC} - w_{LB}}{w_{LB} - w_{GS}} (C_{GS}^{(LB)} - V_{LB}^{(GS)}) - V_{LB}^{(GS)} > 0 \tag{25}$$

the left-hand side of (24) is decreasing in $p_{LB} \in (0, 1)$ and thus, by (19), it is decreasing in R (for such R for which $p_{LB} \in (0, 1)$). Consequently, the necessary condition for (24) to be satisfied for some $R \in (R_{\min}, R_{\max})$ is that it is satisfied for $R = R_{\max}$, i.e. for $p_{LB} = 1$. Thus, we need

$$\frac{w_{SC} - w_{LB}}{w_{LB} - w_{GS}} \big(-C_{GS}^{(LB)} + (E_{GS} - E_{LB})\big) + (E_{SC} - E_{LB}) < C_{SC}^{(H)}. \tag{26}$$

The sufficient condition for non-invadability of a proper mixture of GS and LB is when (24) is satisfied for $p_{LB} = 0$, i.e. when

$$\frac{w_{SC} - w_{LB}}{w_{LB} - w_{GS}} (E_{GS} - E_{LB} - V_{LB}^{(GS)}) + (E_{SC} - E_{LB}) < C_{SC}^{(H)} + V_{LB}^{(GS)}. \tag{27}$$

It is clear from both the necessary condition, (26), and the sufficient condition, (27), that large $C_{SC}^{(H)}$ prevents the invasions of SC, while large E_{SC} and large $(w_{SC} - w_{LB})/(w_{LB} - w_{GS})$ helps the invasion. Also, it follows from above that SC is more likely to invade when $R \approx R_{\min}$.

3.3 Non-invadability by SC

In the previous section we considered conditions under which the native fish population is in a stable mixture that SC cannot invade. However, in order to find a measure to control the invasion of SC, we will now consider scenarios under which the native fish population is not in a mixture but rather a uniform population of a single species (i.e., when $R < R_{\min}$ or $R > R_{\max}$).

When $R < R_{\min}$, the population of native fish will converge to an equilibrium of $p_{LB} = 0$, i.e. GS only population. In such a population we have, by (4) and (2),

$$\mathscr{E}_{GS} = R + E_{GS}, \tag{28}$$

$$\mathscr{E}_{SC} = \frac{w_{SC}}{w_{GS}} R + E_{SC} - C_{SC}^{(H)}. \tag{29}$$

Thus, an advantage of native fish is given by

$$\mathscr{E}_{GS} - \mathscr{E}_{SC} = -\left(\frac{w_{SC}}{w_{GS}} - 1\right) R - (E_{SC} - E_{GS}) + C_{SC}^{(H)} \tag{30}$$

which means that the advantage decreases with increasing R. Consequently, the native fish advantage is highest when $R \approx 0$.

Similarly, when $R > R_{\max}$, then $p_{LB} = 1$ (i.e., LB only population) is a stable state and thus by (3) and (2) we get that the advantage of native fish is given by

$$\mathscr{E}_{LB} - \mathscr{E}_{SC} = -\left(\frac{w_{SC}}{w_{LB}} - 1\right) R - (E_{SC} - E_{LB}) + C_{SC}^{(H)} \tag{31}$$

which means that the advantage decreases with increasing R. Consequently, the native fish advantage is highest when $R \approx R_{\max}$.

Note that the advantage of native fish species over the invasive species is highest when $R \approx 0$ and the necessary condition for the native species to ever have an advantage is that (30) is satisfied for $R = 0$, i.e. we need

$$C_{SC}^{(H)} > E_{SC} - E_{GS}. \tag{32}$$

Figure 2 illustrates the results of this and the previous section.

4 Conclusions and Discussions

We have built and analyzed a game-theoretical model of a *Silver Carp* invasion into the Mississippi river's population of *Gizzard Shad* and *Largemouth Bass*. The first species represent an invasive Asian Carp species and the two latter

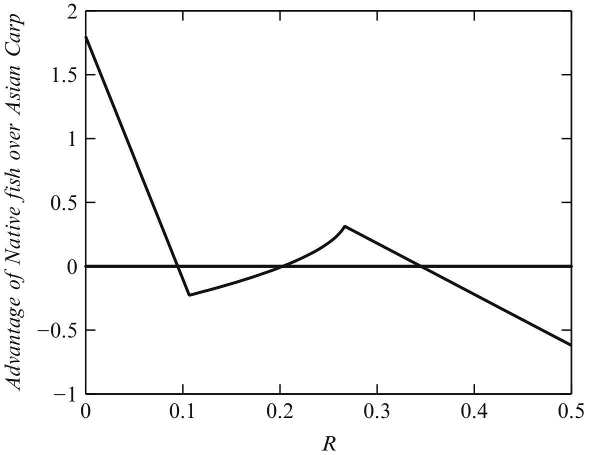


Fig. 2 The fitness advantage of native species over the *Silver Carp* species ($\mathcal{E}_{GS} - \mathcal{E}_{SC}$ for $R < R_{\min}$ and $\mathcal{E}_{LB} - \mathcal{E}_{SC}$ otherwise), as $p_{SC} = 0$, GS and LB is in equilibrium and R varies. The parameter values are $w_{SC} = 1, w_{LB} = 0.2$, $w_{GS} = 0.05$, $E_{SC} = 4.2$, $E_{LB} = 0.08$, $E_{GS} = 0.5$, $V_{LB}^{(GS)} = 0.1$, $C_{GS}^{(LB)} = 0.22$, $C_{SC}^{(H)} = 5.5$

species represent general predator–prey native species in the Mississippi river. Our model took into an account the fact that *Silver Carp* eats large amounts of algae and other microorganisms and thus depletes the resources for native species. We also incorporated the *Silver Carp*'s high (relative to native species) reproduction rate. In the interest of the simplicity of our model, several factors were either completely neglected or simplified. For example, we did not explicitly incorporate the harvesting of native species by humans (or other predators) because such a cost could up to a large extent be included by lowering the reproduction rate. Also, an explicit formulation of the dynamics would be needed to study the conditions under which the native species can persist in the population even after the Asian Carp successfully invaded it.

We were mostly concerned with the prevention of the Asian Carp invasion, so we did not specifically model the spreading of Asian Carp to other places (which happens typically only when the invasive species reaches relatively large density [6]). Consequently, we did not have to incorporate the cost of such spreading in too much detail. However, such details are needed in order to estimate how increasing the cost of spreading (for example, by building electric barriers in the river) may help prevent the Asian Carp invasion.

Despite our model being relatively simple, it still provides enough insights. The analysis of our model shows that (1) the proportion of native predators positively correlates with the available resources in the river and (2) for a healthy population of native species to exist, one needs to have the resources within a reasonable range. If the resources drop below a certain threshold, the fish population tends to *Gizzard Shad* only (i.e., the predators go extinct) while if the resources grow above another threshold, the population tends to *Largemouth Bass* only (the prey species goes extinct).

We have shown conditions under which Asian Carp cannot invade native fish populations. Due to the high reproductive advantage over the native species, one of the best ways to stop Asian Carp is to inflict cost to them (for example, by selectively harvesting them). Also, perhaps surprisingly, within the range of resource level when both native fish (predator and the prey) can coexist, increasing the resource level helps native fish to gain some advantage over the invasive species. However, increasing the resource levels comes with the increase of predatory fish population, potentially making the prey fish population endangered.

We have also discovered a counterintuitive measure to control Asian Carp invasion. By artificially decreasing the resource levels so much that not only predatory species go extinct but also the native prey species start to suffer, one can substantially increase the native fish advantage over the invasive species. Such measure is not sustainable in a long run and implementing it would probably require supplying native fish species into the river to prevent their complete elimination. However, when coupled with an increased effort to eliminate the Asian Carp, it may be the most effective way to stop the invasion.

Acknowledgements This research was supported by MAA's NREUP program funded by NSA (grant H98230-13-1-0270) and NSF (grant DMS-1156582). The authors also would like to thank Dr. Tsvetanka Sendova for her support and guidance during the NREUP program.

References

1. Broom M, Rychtář J (2013) Game-theoretical models in biology, vol 48. CRC Press, Boca Raton
2. Chick JH, Pegg MA (2001) Invasive carp in the Mississippi River basin. Science 292(5525):2250–2251
3. Congress (2012) Stop invasive species act. http://beta.congress.gov/bill/112th-congress/senate-bill/2317 [Online]. Accessed 20 April 2014
4. Hofbauer J, Sigmund K (1988) The theory of evolution and dynamical systems. Cambridge University Press, Cambridge
5. White House (2013) Obama administration releases 2013 Asian carp control strategy framework. Council on environmental quality. http://www.whitehouse.gov/administration/eop/ceq/Press_Releases/July_24_2013 [Online]. Accessed 11 March 2014
6. Koel TM, Irons KS, Ratcliff EN (2000) Asian carp invasion of the upper Mississippi River system. US Department of the Interior, US Geological Survey, Upper Midwest Environmental Sciences Center
7. City of Chicago and Great Lakes Fishery Commission (2012) Fy2012 Asian carp control strategy framework. http://asiancarp.us/documents/2012Framework.pdf [Online]. Accessed 20 April 2014
8. Department of Natural Resources. Largemouth bass. http://www.dnr.state.md.us/fisheries/fishfacts/lgmouthbass.asp [Online]. Accessed 20 April 2014
9. National Park Service (2014) Asian carp overview
10. U.S. Geological Survey USGC (2012) Asian carp. http://www.umesc.usgs.gov/invasive_species/asian_carp.html [Online]. Accessed 11 March 2014
11. Wanner GA, Klumb RA (2009) Length–weight relationships for three asian carp species in the Missouri river. J Freshwater Ecol 24(3):489–495

A Comparison of Multiple Genome-Wide Recombination Maps in *Apis mellifera*

Caitlin Ross, Dominick DeFelice, Greg Hunt, Kate Ihle, and Olav Rueppell

1 Introduction

During the production of gametes (meiosis), regions of homologous chromosomes disconnect and switch places through a process called homologous recombination. Although the explanations for this process are not precisely known, it is a ubiquitous process during meiosis [17]. Homologous recombination serves to stabilize chromosome pairs during meiosis, which provides a mechanistic explanation for meiotic recombination [2]. However, adaptive explanations gain support from findings that recombination rate varies among and between species [16, 20]. Ultimately, it is essential to understand what causes variation in local recombination rates and to understand at what scale DNA sequence motif evolution drives recombination rates [11, 13] and may cause correlation in local rates of recombination between recombination maps [20].

C. Ross
Department of Computer Sciences, The University of North Carolina at Greensboro,
Greensboro, NC 27402, USA

D. DeFelice
Department of Biology, The University of North Carolina at Greensboro,
Greensboro, NC 27402, USA

G. Hunt
Department of Entomology, Purdue University, West Lafayette, IN 47907, USA

K. Ihle
School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA

O. Rueppell (✉)
Department of Biology, 312 Eberhart Building, The University of North Carolina at Greensboro,
321 McIver Street, Greensboro, NC 27402, USA
e-mail: olav_rueppell@uncg.edu; orueppell@gmail.com

J. Rychtář et al. (eds.), *Collaborative Mathematics and Statistics Research*, Springer
Proceedings in Mathematics & Statistics 109, DOI 10.1007/978-3-319-11125-4_10

Among multicellular animals, the genome-wide rate of meiotic recombination, measured in cM/Mb (genetic distance divided by physical distance), is highest in the honey bee, *Apis mellifera*, and other social insects [3]. Currently, no ultimate or proximate hypothesis sufficiently accounts for these extra-ordinarily high recombination rates [19, 25]. Ultimately, recombination is a relatively weak force to increase intra-colonial genetic variation [19] and proximately, the low GC content of the honey bee genome [25] contrasts with the finding that GC content is usually positively correlated with recombination rate [9].

Several studies have analyzed the patterns of recombination in the honey bee genome at coarse [3] or fine [12] scales, using singular mapping populations. However, recombination patterns vary intra-specifically [4] and theoretical reasons suggest recombination rate in honey bees may be even more variable than in other species [24]. In this study, we compare recombination rates calculated from eight different genetic maps of the honey bee using different scales of analysis. Specifically, we tested the prediction that similarity among maps decreases with increasing resolution of analysis, based on the proximate hypothesis that specific sequence motifs drive the patterns or recombination. We assume that these sequence motifs change at a more local scale than the position or function of genes. Consequently, we predict less scale-dependence if adaptive reasons cause intra-genomic variation in recombination rates in the honey bee. Based on the same rationale, we predict that the phylogenetic distance (=coalescence time) between mapping populations is positively correlated to their recombination differences.

2 Methods

Data from eight existing, genome-wide recombination maps of the honey bee were gathered. These maps were constructed in different laboratories using different genetic markers and marker densities (Table 1). Two map pairs (R3/R5 and LBC/HBC) were constructed from sister queens, although the two pairs were not related to each other. The LBC/HBC pair represented reciprocal backcrosses from strains of honey bees that had been artificially selected based on pollen hoarding behavior for >15 generations [18], while the R3/R5 maps represent the recombination events of hybrid queens between Africanized and European stock in Arizona [8]. The JH map was also constructed from a backcross derived from the artificially selected pollen hoarding strains but the mapping population was derived several generations and at least one outcrossing event [7] after the LBC/HBC pair. The VSH map was developed using a queen from a line of bees selectively bred for a behavior associated with mite resistance, derived from commercial stock in the USA and maintained by the USDA [23]. The Grooming map was constructed from commercial stock in the USA selected based on another behavioral phenotype associated with mite resistance [1]. The offspring of two commercial queens from France were used to create the Solignac map [21]. These markers were constructed with different marker technologies and numbers (Table 1).

Table 1 Eight compared recombination maps with their basic information

Name	Marker type	# of markers	Total recombinational length (cM)	Citation
Solignac	Microsatellite	2,008	4,000	[21]
LBC	Microsatellite and SNP (Sequenom MALDI-TOF)	221	3,933	[18]
HBC	Microsatellite and SNP (Sequenom MALDI-TOF)	231	4,135	[18]
R3	Microsatellite and SNP (Sequenom MALDI-TOF)	235	4,747	[8]
R5	Microsatellite and SNP (Sequenom MALDI-TOF)	231	4,319	[8]
VSH	SNP (Illumina chip)	1,340	5,340	[23]
Grooming	SNP (Illumina chip)	1,313	5,403	[1]
JH	SNP (Floragenex RAD-tag sequencing)	1,125	5,696	[10]

Genetic sequences were obtained for all loci in each of the eight maps. Using these sequences, we ran BLAST-n searches to identify the physical location of each marker in the honey bee genome (Amel 4.5). Other data, such as the physical length of chromosomes and gene content was also downloaded from Amel 4.5, using the NCBI website (http://www.ncbi.nlm.nih.gov/genome/?term=apis%20mellifera%20genome). The physical locations of the markers were used in combination with the recombinational inter-marker distances to calculate recombination rates in centimorgans per megabase (cM/Mb) between each consecutive pair of genetic markers in each of the eight data sets. These sets of recombination rates were examined for inconsistencies between the orders of physical and genetic positions. When the marker order in the genetic map was inverted relative to the physical map, we excluded the respective intervals from the analysis. The recombination rates that were consistent for physical and genetic position were then used to calculate recombination rates for each analysis window in the genome. Starting at the beginning of each chromosome, we used window sizes of 50, 100, 250, 500, and 1,000 kbp.

Weighted averages were calculated when multiple inter-marker intervals spanned a window. Any recombination rates greater than 2,000 cM/Mb were considered to be unrealistic and were omitted as outliers. The eight data sets were then compared using SPSS 21.0 (IBM) for each of the five different scales. We studied the effects of map, chromosome, and analysis unit on average recombination rate estimates using ANOVA. We also determined pairwise correlations between each pair of maps for each window size. The chromosomal averages of recombination rate were also computed and a hierarchical cluster analysis of the eight maps was performed based on these values to evaluate overall similarity of the maps with regard to recombination rate.

3 Results

Across all eight maps, we analyzed 26,664 50 kbp, 13,738 100 kbp, 5,836 250 kbp, 3,096 500 kbp, and 1,638 1,000 kbp windows. For the three smaller window sizes, the local recombination rates were significantly different among chromosomes and maps, with significant interactions but the effect of chromosome decreased with increasing window size (Table 2). The correlation of local recombination rates among maps was not significantly affected by window size (Fig. 1). This result held true when individual chromosomes were analyzed. Regardless of window size, some chromosomes exhibited stronger correlations across the eight genetic maps than others, as shown for 250 kbp sized windows (Fig. 2). Average correlation coefficients of the chromosome-specific recombination rates were not related to chromosomal size, gene and transcript number, or their density. The cluster analysis indicated methodological influences on the similarity of chromosome-specific recombination rates with the most fundamental separation between clusters based on low versus high marker number (Fig. 3). In addition, the VSH and the Grooming map clustered most closely and formed a larger cluster with the JH and Solignac map, regardless of the origin of the mapping populations.

Table 2 Factorial ANOVAs indicated significant differences among chromosomes and maps at all scales

Scale (kbp)	Factor	ANOVA result
50	Map	$F_{(7,26536)} = 99.9, p < 0.001$
	Chromosome	$F_{(15,26536)} = 11.1, p < 0.001$
	Map × Chromosome	$F_{(105,26536)} = 6.0, p < 0.001$
100	Map	$F_{(7,13610)} = 57.3, p < 0.001$
	Chromosome	$F_{(15,13610)} = 6.5, p < 0.001$
	Map × Chromosome	$F_{(105,13610)} = 3.7, p < 0.001$
250	Map	$F_{(7,5708)} = 45.0, p < 0.001$
	Chromosome	$F_{(15,5708)} = 2.8, p < 0.001$
	Map × Chromosome	$F_{(105,5708)} = 1.6, p < 0.001$
500	Map	$F_{(7,2968)} = 36.1, p < 0.001$
	Chromosome	$F_{(15,2968)} = 1.5, p = 0.090$
	Map × Chromosome	$F_{(105,2968)} = 1.3, p = 0.026$
1,000	Map	$F_{(7,26536)} = 44.8, p < 0.001$
	Chromosome	$F_{(15,26536)} = 1.1, p = 0.308$
	Map × Chromosome	$F_{(105,26536)} = 0.9, p = 0.838$

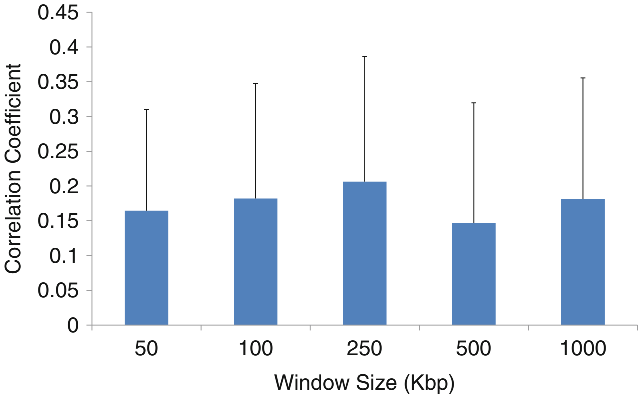


Fig. 1 The correlation coefficients of all pairwise comparisons of recombination rates for the eight genetic maps averaged for each window size over all chromosomes (means are given with standard deviations). Scale (window size) did not affect the correlation among maps

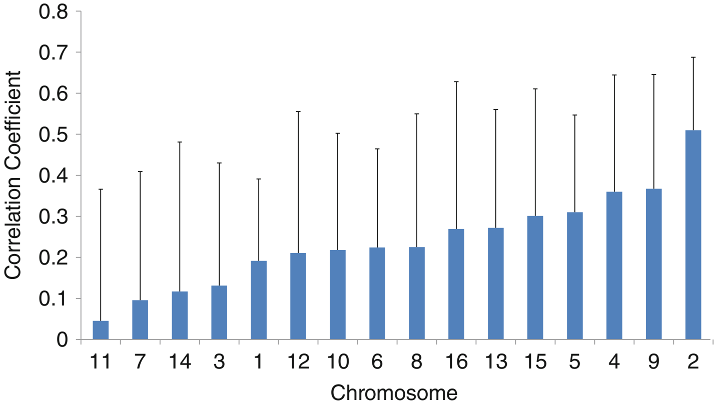


Fig. 2 The average pairwise correlations between maps differed significantly among chromosomes, regardless of the scale of the analysis. Mean correlation coefficients (±standard deviation) of all pairwise comparisons of recombination rates for the eight genetic maps are shown here for each chromosome using a 250 kbp analysis window

4 Discussion

Our comparison of the multiple recombination maps is the first to shed light on the intra-specific variation of meiotic recombination in *A. mellifera*. The correlations among multiple recombination maps in this species were only modest, particularly when compared to an inter-specific analysis in *Drosophila* [22]. For *A. mellifera* with its high overall recombination rate, low correlations are predicted, based on the current model of recombination that postulates a self-destructive nature of the DNA motifs that mediate recombination [15, 24].

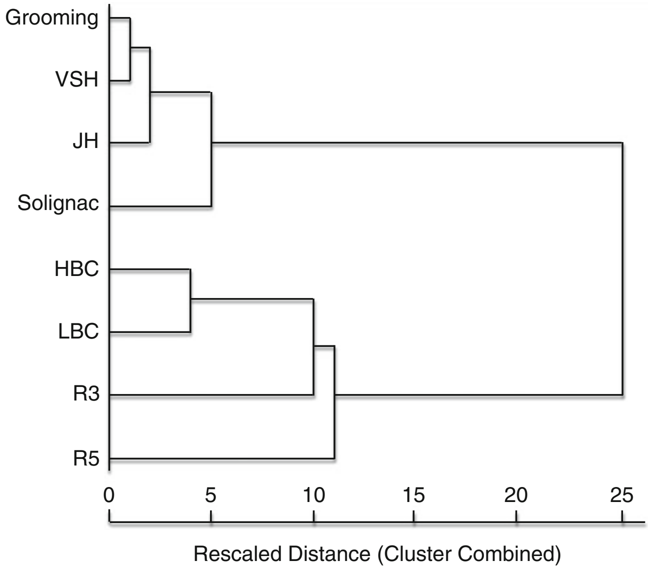


Fig. 3 Based on chromosomal averages of recombination rate for each of the eight maps, these maps were hierarchically clustered. The resulting dendrogram indicates that methodological similarities, especially marker number, influenced the correlation among recombinational estimates at the chromosome scale more than coalescence patterns

Previous studies have found that the correlation among recombination maps depends on the scale of analysis in mammals and *Drosophila*, with correlations typically increasing with increasing scale [11, 14, 22]. The five scales of analysis in our study varied twentyfold, which may have been insufficient to discern significant differences in the correlations among recombination maps. Furthermore, our scaling analysis ultimately depended on the inter-marker distances, which was limiting for the finer scales for some of the linkage maps. Nevertheless, our result that the correlations among recombination rates in different mapping populations are scale independent favors an adaptive explanation of intra-specific variation of recombination rate in *A. mellifera*. Similarly, the lack of phylogenetic signal in the overall similarity pattern of recombination rates suggests a negligible effect of local sequence motifs, that may present a mechanistic, non-adaptive explanation for intra-specific variation in the recombination rate of *A. mellifera*.

Our analyses indicated that the average recombination rate estimates differed when different sizes of analysis windows were used. Theoretically, the averaging over differently sized intervals should yield similar results. However, in practice the differences can be explained by the inclusion of statistical outliers in larger windows due to the averaging effect. In addition, missing data may bias the average recombination rate estimates of small and large intervals differently.

The variable conservation of recombination rates among different chromosomes is intriguing but none of the investigated chromosomal features provided an explanation why the average correlation coefficients varied from 0.05 for chromosome 11 over tenfold to 0.51 for chromosome 2. Comparative studies of genomic recombination rates are still only beginning to emerge, despite their potential richness in information [4]. Despite clear evidence that chromosomes differ in average recombination rate within many species, which we also found here for *A. mellifera*, the question whether the conservation of recombinational landscapes is dependent on chromosome identity within a genome needs to be addressed more to understand the evolution of recombination [5]. To our knowledge, the only known species that have been analyzed this way are humans [6] and *Drosophila melanogaster* [4]. With the construction of denser linkage maps, the exceptionally high rate of recombination of the honey bee may yield a particularly detailed view of factors that influence local recombination rates and their preservation in the future.

Acknowledgements We would like to thank the members of the UNCG Social Insect lab and the Math-Bio working group. The work was funded by the National Science Foundation (grants DMS 0850465 and DBI 0926288) and additionally supported by the National Institutes of Health (NIGMS grant R15GM102753).

References

1. Arechavaleta-Velasco ME, Alcala-Escamilla K, Robles-Rios C, Tsuruda JM, Hunt GJ (2012) Fine-scale linkage mapping reveals a small set of candidate genes influencing honey bee grooming behavior in response to Varroa mites. PLoS One 7:e47269
2. Baker BS, Carpenter ATC, Esposito MS, Esposito RE, Sandler L (1976) The genetic control of meiosis. Ann Rev Genet 10:53–134
3. Beye M, Gattermeier I, Hasselmann M, Gempe T, Schioett M, Baines JF, Schlipalius D, Mougel F, Emore C, Rueppell O, Sirvio A, Guzman-Novoa E, Hunt G, Solignac M, Page RE (2006) Exceptionally high levels of recombination across the honey bee genome. Genome Res 16:1339–1344
4. Comeron JM, Ratnappan R, Bailin S (2012) The many landscapes of recombination in *Drosophila melanogaster*. PLoS Genet 8:e1002905
5. Coop G, Przeworski M (2007) An evolutionary view of human recombination. Nat Rev Genet 8:23–34
6. Coop G, Wen XQ, Ober C, Pritchard JK, Przeworski M (2008) High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. Science 319:1395–1398
7. Dixon L, Kuster R, Rueppell O (2014) Reproduction, social behavior, and aging trajectories in honeybee workers. AGE 36:89–101
8. Graham AM, Munday MD, Kaftanoglu O, Page RE Jr, Amdam GV, Rueppell O (2011) Support for the reproductive ground plan hypothesis of social evolution and major QTL for ovary traits of Africanized worker honey bees (*Apis mellifera L.*). BMC Evol Biol 11:95
9. Hey J, Kliman RM (2002) Interactions between natural selection, recombination and gene density in the genes of drosophila. Genetics 160:595–608
10. Ihle KE, Rueppell O, Page RE, Amdam GV (unpublished) QTL for ovary size and juvenile hormone response to Vg-RNAi knockdown. J Hered
11. Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ (2004) Comparative recombination rates in the rat, mouse, and human genomes. Genome Res 14:528–538

12. Kent CF, Minaei S, Harpur BA, Zayed A (2012) Recombination is associated with the evolution of genome structure and worker behavior in honey bees. Proc Natl Acad Sci USA 109:18012–18017
13. Kulathinal RJ, Bennettt SM, Fitzpatrick CL, Noor MAF (2008) Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. Proc Natl Acad Sci USA 105:10051–10056
14. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321–324
15. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P (2010) Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science 327:876–879
16. Nachman MW (2002) Variation in recombination rate across the genome: evidence and implications. Curr Opin Genet Dev 12:657–663
17. Otto SP, Lenormand T (2002) Resolving the paradox of sex and recombination. Nat Rev Genet 3:252–261
18. Rueppell O, Metheny JD, Linksvayer TA, Fondrk MK, Page RE Jr, Amdam GV (2011) Genetic architecture of ovary size and asymmetry in European honeybee workers. Heredity 106:894–903
19. Rueppell O, Meier S, Deutsch R (2012) Multiple mating but not recombination causes quantitative increase in offspring genetic diversity for varying genetic architectures. PLoS One 7:e47220
20. Smukowski CS, Noor MAF (2011) Recombination rate variation in closely related species. Heredity 107:496–508
21. Solignac M, Mougel F, Vautrin D, Monnerot M, Cornuet JM (2007) A third-generation microsatellite-based linkage map of the honey bee, *Apis mellifera*, and its comparison with the sequence-based physical map. Genome Biol 8:R66
22. Stevison LS, Noor MAF (2010) Genetic and evolutionary correlates of fine-scale recombination rate variation in *Drosophila persimilis*. J Mol Evol 71:332–345
23. Tsuruda JM, Harris JW, Bourgeois L, Danka RG, Hunt GJ (2012) High-resolution linkage analyses to identify genes that influence Varroa Sensitive Hygiene behavior in honey bees. PLoS One 7:e48276
24. Ubeda F, Wilkins JF (2011) The Red Queen theory of recombination hotspots. J Evol Biol 24:541–553
25. Weinstock GM, Robinson GE, Gibbs RA, Worley KC, Evans JD, Maleszka R, Robertson HM, Weaver DB, Beye M, Bork P, Elsik CG, Hartfelder K, Hunt GJ, Zdobnov EM, Amdam GV, Bitondi MMG, Collins AM, Cristino AS, Lattorff HMG, Lobo CH, Moritz RFA, Nunes FMF, Page RE, Simoes ZLP, Wheeler D, Carninci P, Fukuda S, Hayashizaki Y, Kai C, Kawai J, Sakazume N, Sasaki D, Tagami M, Albert S, Baggerman G, Beggs KT, Bloch G, Cazzamali G, Cohen M, Drapeau MD, Eisenhardt D, Emore C, Ewing MA, Fahrbach SE, Foret S, Grimmelikhuijzen CJP, Hauser F, Hummon AB, Huybrechts J, Jones AK, Kadowaki T, Kaplan N, Kucharski R, Leboulle G, Linial M, Littleton JT, Mercer AR, Richmond TA, Rodriguez-Zas SL, Rubin EB, Sattelle DB, Schlipalius D, Schoofs L, Shemesh Y, Sweedler JV, Velarde R, Verleyen P, Vierstraete E, Williamson MR, Ament SA, Brown SJ, Corona M, Dearden PK, Dunn WA, Elekonich MM, Fujiyuki T, Gattermeier I, Gempe T, Hasselmann M, Kadowaki T, Kage E, Kamikouchi A, Kubo T, Kucharski R, Kunieda T, Lorenzen MD, Milshina NV, Morioka M, Ohashi K, Overbeek R, Ross CA, Schioett M, Shippy T, Takeuchi H, Toth AL, Willis JH, Wilson MJ, Gordon KHJ, Letunic I, Hackett K et al (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443:931–949

Part 3

Constructing Splitting Fields of Polynomials over Local Fields

Jonathan Milstead, Sebastian Pauli, and Brian Sinclair

1 Introduction

Let K be a local field. We present an algorithm that given a polynomial $\Phi \in \mathscr{O}_{\mathsf{K}}[x]$ computes the splitting field L of Φ, that is $\mathsf{K}(\theta_1, \dots, \theta_N)$ where $\theta_1, \dots, \theta_N$ are the roots of Φ.

Our algorithm is a variation of an OM algorithm [2], that is specialized to the computation of splitting fields. OM algorithms are named after Ore–MacLane or Okutsu–Montes and are algorithms that compute the OM invariants of a polynomial Φ [7, 14] which can be used to compute the factorization of Φ, an integral basis of the fields generated by the irreducible factors of Φ, their inertia degree and ramification index, and the decomposition of primes in the maximal order of the global field generated by Φ. Several algorithms have been developed for these tasks, for example by MacLane [11] (prime decomposition), Ford and Zassenhaus [4] (integral bases), Okutsu [14] (integral bases), and Montes [12] (prime decomposition). All of them compute the OM invariants more or less explicitly. Our algorithm is based on Montes' algorithm [8, 12, 17].

Our method for computing splitting fields uses the information about the extension generated by the roots of Φ computed in each iteration of the OM algorithm to construct subfields of L of Φ until the splitting field is obtained. In particular we extend our approximation L to the splitting field as soon as we find inertia or that the splitting field contains a certain tamely ramified subextension. We generate wildly ramified subfields of L as soon as we have found an irreducible factor of Φ generating such an extension. In our representation of the algorithm we

J. Milstead • S. Pauli (✉) • B. Sinclair
Department of Mathematics and Statistics, University of North Carolina Greensboro, Greensboro, NC 27402, USA
e-mail: jmmilste@uncg.edu; s_pauli@uncg.edu; basincla@uncg.edu

J. Rychtář et al. (eds.), *Collaborative Mathematics and Statistics Research*, Springer Proceedings in Mathematics & Statistics 109, DOI 10.1007/978-3-319-11125-4_11

follow the approach in [18]. An application of our algorithm can be found in [1]. It can also be modified into a root separation algorithm.

A method, similar to ours, is available in Magma [3]. It makes multiple calls to a variant of the Round 4 algorithm, but does not make use of all the information computed internally. The Round 4 algorithm is an OM algorithm that is less efficient than the algorithms based on Montes' methods [12]. The implementation of the Round 4 algorithm (a combination of the algorithms described in [5] and [16]) in Magma returns the factorization of a polynomial over a local field and in addition the extensions generated by the irreducible factors of the polynomial. In the computation of splitting fields, an initial call to Round 4 is used to generate the unramified extension whose degree is the least common multiple of the inertia degrees of the extensions generated by the irreducible factors of Φ. Then Φ is considered over the extended field. After each subsequent call to Round 4, the field is extended using one of the irreducible factors of Φ over the current approximation to the splitting field until Φ splits into linear factors.

1.1 Overview

Section 2 contains a description of situations when factorizations of polynomials can be derived. Although we do not use these methods directly (they are used in the Round 4 algorithm), they make understanding the algorithm easier. In Sect. 3 we give some results about tamely ramified extensions and composites of tamely ramified extensions. The general strategy of the splitting field algorithm is outlined in Sect. 4 with technical details provided in Sects. 5 and 6. We present the splitting field algorithm and some auxiliary algorithms in Sect. 7.

1.2 Notation

Let K be a local field, that is a field complete with respect to a non-archimedian exponential valuation v with finite residue class field $\underline{\mathsf{K}} \cong \mathbb{F}_q$ of characteristic p. We denote the multiplicative group of K by $\mathsf{K}^\times$. Let $\mathcal{O}_\mathsf{K}$ be the valuation ring of K and let $\Phi(x) \in \mathcal{O}_\mathsf{K}[x]$. For our purposes, $v = v_\mathsf{K}$ is normalized such that $v_\mathsf{K}(\pi_\mathsf{K}) = 1$ where $\pi = \pi_\mathsf{K}$ is a uniformizing element in $\mathcal{O}_\mathsf{K}$. For $\gamma \in \mathcal{O}_\mathsf{K}$ we denote by $\underline{\gamma}$ the class $\gamma + (\pi)$ in $\underline{\mathsf{K}}$. The unique extension of v to an algebraic closure $\overline{\mathsf{K}}$ of K (or to any intermediate field) is also denoted v. If L/K is a finite extension, then v_L denotes the valuation that is normalized such that $v_\mathsf{L}(\pi_\mathsf{L}) = 1$ where π_L is a uniformizing element of L.

Definition 1. For $\gamma \in \overline{\mathsf{K}}^\times$ and $\delta \in \overline{\mathsf{K}}^\times$ we write $\gamma \sim \delta$ if

$$v(\gamma - \delta) > v(\gamma)$$

and impose the supplementary condition $0 \sim 0$. For $\varphi(x) = \sum_{i=0}^{n} c_i x^i$ and $\psi(x) = \sum_{i=0}^{n} b_i x^i$ in $\overline{\mathsf{K}}[x]$ we write $\varphi \sim \psi$ if

$$\min{}_{0 \le i \le n}\ v(c_i - d_i)\ >\ \min{}_{0 \le i \le n}\ v(c_i).$$

It follows immediately that the relation $\sim$ is symmetric, transitive, and reflexive. Let L be a finite extension of K with uniformizing element π_{L}. Two elements $\gamma = \gamma_0 {\pi_{\mathsf{L}}}^u \in \mathsf{L}$ and $\delta = \delta_0 {\pi_{\mathsf{L}}}^w \in \mathsf{L}$ with $v(\gamma_0) = v(\delta_0) = 0$ are equivalent with respect to $\sim$ if and only if $u = w$ and $\gamma_0 \equiv \delta_0 \bmod (\pi_{\mathsf{L}})$.

2 Hensel Lifting and Newton Polygons

Hensel lifting yields a factorization of polynomials over local fields in certain cases and Newton polygons give valuable information about the roots of polynomials. We show how these two tools can be used to obtain proper factorizations in more general cases.

Theorem 1 (Hensel's Lemma). *Let $\Phi \in \mathcal{O}_{\mathsf{K}}[x]$ be monic. If $\Phi \equiv \phi_1\phi_2 \bmod (\pi)$ where ϕ_1 and ϕ_2 are coprime modulo π, then there is a factorization $\Phi = \Phi_1\Phi_2$ with $\Phi_1 \equiv \phi_1 \bmod (\pi)$ and $\Phi_2 \equiv \phi_2 \bmod (\pi)$.*

For an example of an efficient Hensel lifting algorithm that lifts a factorization modulo (π) to a factorization modulo $(\pi)^s$ for any given s, see [19]. We can also obtain an approximation to a factorization of Φ if Hensel lifting can be applied to the characteristic polynomial of an element $\phi + (\Phi)$ in $\mathcal{O}_{\mathsf{K}}[x]/(\Phi)$.

Definition 2. Let $\Phi(x) = \prod_{j=1}^{N}(x - \theta_j) \in \mathcal{O}_{\mathsf{K}}[x]$. For $\phi \in \mathsf{K}[x]$ we define

$$\chi_\phi(y) := \prod_{i=1}^{N}(y - \phi(\theta_i)) = \mathrm{res}_x(\Phi(x), y - \phi(x)) \in \mathsf{K}[y].$$

Proposition 1. *Let $\gamma \in \mathsf{K}[x]$ with $\chi_\gamma \in \mathcal{O}_{\mathsf{K}}[y]$. If $\underline{\chi}_\gamma$ has at least two distinct irreducible factors, then Φ is reducible in $\mathcal{O}_{\mathsf{K}}[x]$.*

Proof. Suppose $\underline{\chi}_\gamma$ has at least two irreducible factors. Then, Hensel's lemma gives relatively prime monic polynomials $\chi_1 \in \mathcal{O}_{\mathsf{K}}[y]$ and $\chi_2 \in \mathcal{O}_{\mathsf{K}}[y]$ with $\chi_1\chi_2 = \chi_\gamma$. Reordering the roots $\theta_1, \ldots, \theta_N$ of Φ if necessary, we may write

$$\chi_1(y) = (y - \gamma(\theta_1)) \cdots (y - \gamma(\theta_r)) \text{ and } \chi_2(y) = (y - \gamma(\theta_{r+1})) \cdots (y - \gamma(\theta_N)),$$

where $1 \le r < N$. It follows that

$$\Phi = \gcd(\Phi, \chi_1(\gamma)) \cdot \gcd(\Phi, \chi_2(\gamma))$$

is a proper factorization of Φ.

Definition 3 (Newton Polygon). Let $\Phi(x) = \sum_{i=0}^{N} c_i x^i$. The lower convex hull of $\{(i, v(c_i)) \mid 0 \leq i \leq N\}$ is the Newton polygon of Φ.

The negatives of the slopes of the segments of the Newton polygon of Φ are the valuations of the roots of Φ. The length of the segment (in x-direction) is the number of roots with this valuation. The negatives of the slopes of the Newton polygon of the characteristic polynomial χ_ϕ of $\phi + (\Phi)$ are the valuations $v(\phi(\theta))$ for the roots θ of Φ. Proposition 1 yields a constructive method for finding a factorization of Φ if χ_ϕ has more than one segment:

Corollary 1. *Let $\phi \in \mathsf{K}[x]$ with $\chi_\phi \in \mathscr{O}_{\mathsf{K}}[y]$. If there are roots θ and θ' of Φ such that $v(\phi(\theta)) \neq v(\phi(\theta'))$, then we can find two proper factors of $\Phi(x)$ over $\mathscr{O}_{\mathsf{K}}[x]$.*

Proof. Let Θ be the set of roots of Φ and let $h/e = \min\{v(\phi(\theta)) \mid \theta \in \Theta\}$. Setting $\gamma := \phi^e/\pi^h$ we get

$$\max\{v(\gamma(\theta)) \mid \theta \in \Theta \text{ and } \gamma(\theta) = 0\} > \min\{v(\gamma(\theta)) \mid \theta \in \Theta \text{ and } \gamma(\theta) = 0\} = 0.$$

Thus Proposition 1 yields a factorization of Φ.

3 Tamely Ramified Extensions

For all $f \in \mathbb{N}$ there is, up to isomorphism, a unique unramified extension of K of degree f. Such an extension can be generated by any monic polynomial of degree f that is irreducible over $\underline{\mathsf{K}}$.

Each totally and tamely ramified extension of degree e can be generated by a polynomial of the form $x^e - \gamma\pi_{\mathsf{K}}$ where $v(\gamma) = 0$. In certain cases we can obtain a generating polynomial of a tamely ramified subextension from a polynomial generating a totally ramified extension.

Proposition 2 ([6, Proposition 2.1]). *Let $n = e_0 p^m$ with $p \nmid e_0$ and let*

$$\phi(x) = x^n + \sum_{i=1}^{n-1} \phi_i x^i + \phi_0 \in \mathscr{O}_{\mathsf{K}}[x]$$

be a polynomial whose Newton polygon is a line of slope $-h/n$, where $\gcd(h, n) = 1$. Let α be a root of ϕ. The maximum tamely ramified subextension M of $\mathsf{L} = \mathsf{K}(\alpha)$ of degree e_0 can be generated by the Eisenstein polynomial $x^{e_0} - (-\psi_0)^b \pi^{e_0 a}$ where a and b are integers such that $ae_0 + bh = 1$ and $\psi_0 \in \mathscr{O}_{\mathsf{K}}[x]$ with $\psi_0 \equiv \phi_0 \bmod (\pi^{h+1})$.

This proposition also yields the standard form for generating polynomials of tamely ramified extensions mentioned above.

Corollary 2. *Let $\phi(x) = \sum_{i=0}^{e} \phi_i x^i \in \mathscr{O}_{\mathsf{K}}[x]$ be an Eisenstein polynomial and assume $p \nmid e$. If $\psi(x) = x^e + \psi_0$ with $\psi_0 \equiv \phi_0 \bmod (\pi^2)$, then the extensions generated by $\phi(x)$ and $\psi(x)$ are isomorphic.*

In our algorithm we need to find the composite of several tamely ramified extension of the same degree.

Proposition 3. *Let $\phi_1(x) = x^e - \gamma_1\pi \in \mathscr{O}_{\mathsf{K}}[y]$ and $\phi_2(x) = x^e - \gamma_2\pi \in \mathscr{O}_{\mathsf{K}}[x]$ with $p \nmid e$ and $v(\gamma_1) = v(\gamma_2) = 0$. Let θ_1 and θ_2 be a roots of ϕ_1 and ϕ_2, respectively. Then the composite of $\mathsf{K}(\theta_1)$ and $\mathsf{K}(\theta_2)$ is the unramified extension of $\mathsf{K}(\theta_1)$ whose degree is the least common multiple f of the degrees of the irreducible factors of $z^e - \underline{\frac{\gamma_2}{\gamma_1}} \in \underline{\mathsf{K}}[z]$.*

Proof. We have

$$\phi_2(\theta_1 x) = (\theta_1 x)^e - \gamma_2\pi = \theta_1^e x^e - \gamma_2\pi = (\gamma_1\pi)x^e - \gamma_2\pi.$$

Dividing by $\gamma_1\pi$ gives $x^e - \frac{\gamma_2}{\gamma_1} = 0$. So the composite of $\mathsf{K}(\theta_1)$ and $\mathsf{K}(\theta_2)$ is the extension of $\mathsf{K}(\theta_1)$ that contains the roots of $\tau(x) = x^e - \frac{\gamma_2}{\gamma_1}$. Since $\gcd(x^e - \frac{\gamma_2}{\gamma_1}, \frac{d}{dx}(x^e - \frac{\gamma_2}{\gamma_1})) = \gcd(x^e - \frac{\gamma_2}{\gamma_1}, ex^{e-1}) = 1$ the polynomial $\underline{\tau}(z) = z^e - \underline{\frac{\gamma_2}{\gamma_1}} \in \underline{K(\theta_1)}[z]$ is squarefree. Denote by f the least common multiple of the degrees of the irreducible factors of $\underline{\tau}$. Then τ splits into linear factors in the unramified extension of $\mathsf{K}(\theta_1)$ of degree f, which is the composite of $\mathsf{K}(\theta_1)$ and $\mathsf{K}(\theta_2)$.

4 Partitions of Zeros and Types

Let $\Phi(x) = x^N + \sum_{i=0}^{N-1} c_i x^i \in \mathscr{O}_{\mathsf{K}}[x]$ be separable and squarefree and let $\Theta_0 = \{\theta_1, \ldots, \theta_N\}$ be the set of zeros of Φ in $\overline{\mathsf{K}}$. We want to find the splitting field of Φ, that is the smallest extension L/K over which Φ splits into linear factors. We successively generate a tower of subfields of L until we have found L.

In this process we partition the set of the zeros of Φ until all partitions contain exactly one zero of Φ. We obtain a tree with root node Θ_0 whose leaves consist of the sets containing exactly one zero of Φ. Every time we find sufficient information about a subfield of L we continue over the corresponding extension.

In our description of the algorithm, we focus on one path from the root node Θ_0 to a leaf. We indicate where branching occurs, thus describing the construction of all root paths in the tree. The nodes of such a root path are subsets of Θ_0, where, if Θ_{u+1} is a child of Θ_u then $\Theta_{u+1} \subseteq \Theta_u$. To each node Θ_u we attach a subfield K_u of the splitting field such that

$$\mathsf{K} \subset \mathsf{K}_1 \subset \cdots \subset \mathsf{K}_u \subset \cdots \subset \mathsf{L}.$$

In our algorithm, we construct these extensions as soon as we find that L has a certain subfield. When we find a divisor f of the inertia degree of L/K, we continue over the unramified extension of degree f. Similarly we construct an unramified extension of degree e when we find $\phi \in \mathscr{O}_{\mathsf{K}_i}$ with $v(\phi(\theta)) = \frac{h}{ep^{\mu}}$ where $\gcd(h, e) = \gcd(e, p) = 1$ and $e \neq 1$. To find generating polynomials of wildly ramified extensions, in addition to the field K_u, we attach a polynomial ϕ_u to the node Θ_u that is an approximation to a polynomial that generates the wildly ramified part of $\mathsf{K}_u(\theta)/\mathsf{K}_u$ for $\theta \in \Theta_u$. When the degree of ϕ_u is the ramification index of $\mathsf{K}_u(\theta)/\mathsf{K}_u$ then ϕ_u is an approximation to a unique factor of Φ that can be lifted to a factor $\hat{\phi}_u$ of Φ over K_u. We continue working over $\mathsf{K}_u[x]/(\hat{\phi}_u)$.

We start the first iteration with a linear monic polynomial $\phi_1 = x + \beta \in \mathscr{O}_{\mathsf{K}}[x]$. The negatives of the slopes of the segments of the Newton polygon of $\Phi(x - \beta)$ are the valuations of the roots of Φ. Then

$$L_1 = \{v(\phi_1(\theta)) \mid \theta \in \Theta_0\}$$

is the set of negatives of the slopes of the segments of the Newton polygon of $\Phi(x - \beta)$. We obtain a partition of Θ_0 into the sets $\{\theta \in \Theta_0 \mid v(\phi_1(\theta)) = \lambda\}$ for $\lambda \in L_1$. By Corollary 1 each of these sets corresponds to a factor of Φ. For some $\lambda_1 \in L_1$ we set

$$\Theta_1^* = \{\theta \in \Theta_0 \mid v(\phi_1(\theta)) = \lambda_1\}. \tag{1}$$

Without computing Θ_1^* explicitly we investigate the extensions generated by the factor $\prod_{\theta \in \Theta_1^*}(x - \theta)$ of Φ further.

Let $\lambda_1 = h_1/d_1$ in lowest terms. Then $v(\phi_1^{d_1}(\theta)/\pi^{h_1}) = 0$ for all $\theta \in \Theta_1^*$. We set

$$R_1 = \left\{\underline{\rho} \in \underline{\mathsf{K}}[z] \mid \underline{\rho} \text{ irreducible and } \underline{\rho}\left(\underline{\phi_1^{d_1}(\theta)/\pi^{h_1}}\right) = \underline{0} \text{ for } \theta \in \Theta_1^*\right\}.$$

Let $f_1 = \operatorname{lcm}\{\deg \underline{\rho} \mid \underline{\rho} \in R_1\}$ and let K_1^* be the unramified extension of K of degree f_1. Then $\mathsf{K}_1^* \subset \mathsf{L}$. Over $\underline{\mathsf{K}}_1^*$ the polynomials in R_1 split into linear factors. Let

$$\Gamma_1 = \{\underline{\phi_1^{d_1}(\theta)/\pi^{h_1}} \in \underline{\mathsf{K}}_1^* \mid \theta \in \Theta_1^*\}$$

be the set of zeros of the polynomials in R_1. By Proposition 1 each $\underline{\gamma} \in \Gamma_1$ corresponds to a factor of Φ over K_1^*. We continue to construct the splitting field of this factor. For some $\underline{\gamma}_1 \in \Gamma_1$ let

$$\Theta_1 = \left\{\theta \in \Theta_1^* \mid \underline{\phi_1^{d_1}(\theta)/\pi^{h_1}} = \underline{\gamma}_1\right\}. \tag{2}$$

If $|\Theta_1| = 1$, then we have reached a leaf of the tree of partitions. Otherwise we write $d_1 = p^{\mu_1} e_1$ with $p \nmid e_1$. If $e_1 = 1$ we set $\mathsf{K}_1 = \mathsf{K}_1^*$.

If $e_1 > 1$, for each $\gamma \in \Gamma_1$ we obtain a tamely ramified extension of K_1^* that is a subfield of L and let K_1 be the composite of these extensions and the unramified extension of K_1^* that contains the e_1th roots of unity.

Over K_1 we have $\phi_1^{p^{\mu_1}}(\theta) \sim \delta$ for some $\delta \in \mathsf{K}_1$ for all $\theta \in \Theta_1$. The ramification index of $\mathsf{K}_1(\theta)/\mathsf{K}_1$ is divisible by p^{μ_1} and we use $\phi_2 = \phi_1^{p^{\mu_1}}(\theta) - \delta$ as a first approximation to an irreducible factor of Φ that generates a wildly ramified extension of K_1. All relevant information from the considerations above can be recovered from the tuple

$$(\phi_1, \lambda_1, \pi^h, y - \underline{\gamma}_1) \in \mathscr{O}_{\mathsf{K}}[x] \times \mathbb{Q} \times \mathsf{K}[x] \times \underline{\mathsf{K}}_1[y].$$

In the second iteration of the algorithm we use ϕ_2 to investigate the subfield of L that contains the roots in Θ_1 further. The set $L_2 = \{v(\phi_2(\theta)) \mid \theta \in \Theta_1\}$ contains the slopes of the Newton polygon of the characteristic polynomial $\phi_2 + \Phi$ over K_1. Each $\lambda \in L_2$ corresponds to a proper factor of Φ (compare Corollary 1). Let $\lambda_2 \in L_2$ be the slope of a segment of the Newton polygon and $\Theta_2^* = \{\theta \in \Theta_1 \mid v(\phi_2(\theta)) = \lambda_2\}$ be the corresponding subset of zeros of Φ. We write $\lambda_2 = h_2/d_2$ in lowest terms and $d_2 = p^{\mu_2^*} e_2$ with $p \nmid e_2$ and set $\mu_2 = \min\{\mu_2^* - \mu_1, 0\}$.

We find $\psi_2 \in \mathsf{K}_1[x]$ with $\deg \psi_2 < \deg \phi_2$ and $v_{\mathsf{K}_1}(\psi_2(\theta)) = h_2/p^{\mu_2^* - \mu_2}$. Let

$$R_2 = \left\{\underline{\rho} \in \underline{\mathsf{K}}_1[z] \mid \underline{\rho} \text{ irreducible and } \underline{\rho}\left(\underline{\phi_2^{e_2 p^{\mu_2}}(\theta)/\psi_2(\theta)}\right) = \underline{0} \text{ for } \theta \in \Theta_2^*\right\},$$

which is the set of irreducible factors of the characteristic polynomial $\phi_2^{e_2 p^{\mu_2}}/\psi_2 + \Phi$ over $\underline{\mathsf{K}}_1$. By Proposition 1 each polynomial in R_2 corresponds to a factor of Φ. Let $f_2 = \operatorname{lcm}\{\deg \underline{\rho} \mid \underline{\rho} \in R_2\}$. The splitting field L of Φ contains the unramified extension K_2^* of degree f_2 of K_1. If $e_2 \neq 1$, L also contains the composite K_2 of certain tamely ramified extensions of K_2^* of degree e_2. Otherwise we set $\mathsf{K}_2 = \mathsf{K}_2^*$. Over K_2 the slope λ_2 of the segment becomes h_2/p^{μ_2}. Let

$$\Gamma_2 = \left\{\underline{\phi_2^{p^{\mu_2}}(\theta)/\psi_2(\theta)} \in \underline{\mathsf{K}}_2^* \mid \theta \in \Theta_2^*\right\}.$$

By Proposition 1 each $\gamma \in \Gamma_2$ corresponds to a factor of Φ over K_2 and a branch of in our tree of partitions. We follow the branch corresponding to some $\underline{\gamma}_2 \in \Gamma_2$ to the node

$$\Theta_2 = \left\{\theta \in \Theta_2^* \mid \underline{\phi_2^{e_2 p^{\mu_2}}(\theta)/\psi_2(\theta)} = \underline{\gamma}_2\right\}.$$

If $|\Theta_2| = 1$, we do not need to investigate this branch of the tree of partitions further.

If $|\Theta_2| = \deg \phi_2 = p^{\mu_1} > 1$, then ϕ_2 is an approximation to an irreducible factor of Φ of degree $\deg \phi_2$ that defines a wildly ramified extension of K_2. We obtain this factor with single factor lifting [9], construct the corresponding wildly ramified extension, and start over at the root node Θ_0 over this extension with a linear polynomial ϕ_1.

Otherwise we use $\phi_3 = \phi_2^{p^{\mu_2}} - \gamma_2 \psi$ as the next approximation to an irreducible factor of Φ that generates a wildly ramified extension of K_2.

All the information obtained in this second iteration of the algorithm is included in or can be recovered from the information in

$$(\phi_2, \lambda_2, \psi_2, z - \underline{\gamma}_2) \in \mathcal{O}_{\mathsf{K}_1}[x] \times \mathbb{Q} \times \mathsf{K}_1[x] \times \underline{\mathsf{K}}_2[z].$$

We inductively continue this process and keep track of the information computed in a sequence of such tuples called *types* (see [8, Definitions 1.21, 1.22 and Section 2.1]). We generalize them insofar as we allow the coefficients of the polynomials in all tuples in a type to be in an extension of K.

Definition 4. Let $\Phi \in \mathcal{O}_{\mathsf{K}}[x]$ and let L/K be a finite. Let $t = (\phi_i, \lambda_i, \psi_1, \underline{\rho}_i)_{1 \le i \le u}$ where

(a) $\phi_1 \in \mathcal{O}_{\mathsf{L}}[x]$ is linear, $\phi_i \in \mathcal{O}_{\mathsf{L}}[x]$ is monic,
(b) $\lambda_i = h_i/d_i \in \mathbb{Q}$ in lowest terms,
(c) $\psi_i \in \mathsf{L}[x]$ with $\deg \psi_i < \deg \phi_i$, and
(d) $\underline{\rho}_i \in \underline{\mathsf{L}}$ irreducible.

We call t an *extended type of Φ over* L, if for all θ in some subset Θ of the set of roots of Φ we have:

(e) $v(\phi_i(\theta)) = \lambda_i$
(f) $v(\psi_i(\theta)) = e\lambda_i$ with $e = \frac{\operatorname{lcm}(d_1,\dots,d_i)}{\operatorname{lcm}(d_1,\dots,d_{i-1})}$,
(g) $\underline{\rho}_i(\underline{\phi_i^e(\theta)/\psi_i(\theta)}) = \underline{0}$, and
(h) $v(\phi_i(\theta)) > v(\phi_{i-1}(\theta))$ and $\deg \phi_i = e \cdot \deg \underline{\rho}_{i-1} \cdot \deg \phi_{i-1}$ for $2 \le i \le u$.

We call $(\phi_i, \lambda_i, \underline{\rho}_i)_{1 \le i \le u}$ a type of Φ over L.

Definition 5. Let $t = (\phi_i, \lambda_i, \psi_1, \underline{\rho}_i)_{1 \le i \le u}$ be an extended type over L. We call t *optimal* if $\deg \phi_{i-1} < \deg \phi_i$ for $2 \le i \le u$ and *complete* if

$$\deg \phi_u = p^{\max\{\mu_i \mid 1 \le i \le u\}} \cdot \deg \underline{\rho}_1 \cdot \dots \cdot \underline{\rho}_u.$$

If $\deg \phi_i$ is a power of p for all $1 \le i \le u$, then we call t a *wild type* of Φ over L.

A type t describes a root path in a tree of partitions of Θ_0. If $t = (\phi_i, \lambda_i, \psi_i, \rho_i)_{1 \le i \le u}$ a wild type over L with a corresponding subset of roots Θ_u, then $\lambda_i = h_i/\pi^{\mu_i^*}$ and $\operatorname{lcm}(p^{m_1}, \dots, p^{m_u}) = p^{\max\{m_1,\dots,m_u\}}$ divides the ramification index of $\mathsf{L}(\theta)$ for $\theta \in \Theta_u$. In our considerations all types are wild and the polynomials $\underline{\rho}$ are linear.

At the end of the uth iteration of our algorithm we construct a polynomial ϕ_{u+1} of degree $p^{\max\{m_1,\dots,m_u\}}$ that is irreducible with $v(\phi_{u+1}(\theta)) > v(\phi_u(\theta))$ for $\theta \in \Theta_u$.

In the following sections we describe methods for constructing ϕ_{u+1}, finding $v(\phi_{u+1}(\theta))$ for all $\theta \in \Theta_u$, constructing ψ_{u+1}, and finding $\underline{\rho}_{u+1}$. We will see that the sets $\Theta_0 \supset \Theta_1 \supset \cdots \supset \Theta_u$ help in the understanding the algorithm, but will never be explicitly needed in actual computations.

If $t = (\phi_i, \lambda_i, \psi_1, \underline{\rho}_i)_{1\le i\le u}$ is an extended type over L and $|\Theta_u| = p^{\max\{m_1,\dots,m_u\}}$, then ϕ_u is an approximation to a unique irreducible factor of Φ over L of degree $p^{\max\{m_1,\dots,m_u\}}$. Using the information in t, this approximation can be lifted to an approximation of arbitrary precision using single factor lifting (see [9]).

5 The First Iteration

We start our description of the construction of the splitting field of $\Phi \in \mathscr{O}_{\mathsf{K}}[x]$ of degree N with the first iteration. We have already gone through these steps in a more conceptual manner in the previous section. As before let $\phi_1 \in \mathscr{O}_{\mathsf{K}}[x]$ be linear and monic, say $\phi_1(x) = x + \beta$, and let Θ_0 denote the set of zeros of Φ in $\overline{\mathsf{K}}$. Although we use the zeros in Θ_0 in our exposition, they are not needed in any of the computations.

5.1 Newton Polygons I

The Newton polygon of $\Phi(x-\beta)$ yields the valuations of the zeros $\theta_1,\dots,\theta_N$ of Φ. Alternatively we can also use the ϕ_1-expansion of Φ:

Definition 6. Let $\Phi \in \mathscr{O}_{\mathsf{K}_u}[x]$ of degree N and $\phi \in \mathscr{O}_{\mathsf{K}_u}[x]$ of degree n be monic polynomials. We call

$$\Phi = \sum_{i=0}^{\lceil N/n\rceil} a_i \phi^i$$

with $\deg(a_i) < n$ the ϕ-expansion of Φ.

If $\Phi = \sum_{i=0}^{\lceil N/\deg\phi_1\rceil} a_i\phi_1^i$ is the ϕ_1-expansion of Φ, then the polynomial $\chi_1(y) = \sum_{i=0}^{\lceil N/\deg\phi_1\rceil} a_i y^i$ has the zeros $\phi_1(\theta)$ where $\theta \in \Theta_0$. We have

$$\chi_1(y) = \Phi(y-\beta) = \chi_{\phi_1}(y) \tag{3}$$

with χ_{ϕ_1} as in Definition 2. The negatives of the slopes of the segments of the Newton polygon of χ_1 are the valuations of $\phi_1(\theta)$ for $\theta \in \Theta_0$. We obtain a partition of Θ_0 into the sets

$$\{\theta \in \Theta \mid v(\phi_1(\theta)) = \lambda\}$$

where λ is the negative of the slope of a segment of the Newton polygon of χ_1. To find the splitting field one continues the algorithm for each of the sets in this partition.

5.2 Residual Polynomial I

Residual (or associated) polynomials were first introduced by Ore [13, 15]. They yield information about the unramified part of the extension generated by the zeros of Φ. Let S be a segment of the Newton Polygon of $\chi_1(y) = \sum_{i=1}^{N} a_i y^i$ (see (3)), let m_1 be the length of S, $(k, v(a_k))$ and $(k + m_1, v(a_{k+m_1}))$ its endpoints, and $\lambda_1 = \frac{v(a_k)-v(a_{k+n})}{m_1} = \frac{h_1}{d_1}$ where $\gcd(h_1, d_1) = 1$ the negative of its slope. If

$$\Theta_1^* = \{\theta \in \Theta_0 \mid v(\phi_1(\theta)) = \lambda_1\},$$

then $|\Theta_1^*| = m_1$. We evaluate χ_1 at $\phi_1(\theta)y$ and obtain a polynomial whose Newton polygon has a horizontal segment of length m_1. For $\theta \in \Theta_1^*$ we consider $\chi_1(\phi_1(\theta)y)$. Using the equivalence relation from Definition 1 we obtain

$$\chi_1(\phi_1(\theta)y) = \sum_{i=0}^{N} a_i(\phi_1(\theta)y)^i \sim \sum_{i=k}^{k+m_1} a_i\phi_1^i(\theta)y^i \sim \sum_{j=0}^{m_1/d_1} a_{jd_1+k}\phi_1^{jd_1+k}(\theta)y^{jd_1+k}$$

The last equivalence holds, because the x-coordinates of the points on the segment of the Newton polygon are of the form $k + jd_1$ with $0 \le j \le m_1/d_1$. Furthermore for $0 \le j \le m_1/d_1$ we have $v(a_{jd_1+k}\phi_1^{jd_1+k}(\theta)) = v(a_k\phi_1^k(\theta))$ and the polynomial is divisible by y^k. Dividing $\chi_1(\phi_1(\theta)y)$ by $\pi^{v(a_k)}\phi_1^k(\theta)y^k$ we obtain a polynomial of degree m_1/d_1 that is equivalent to a polynomial whose leading coefficient and constant coefficient have valuation zero:

$$\frac{\chi_1(\phi_1(\theta)y)}{\pi^{v(a_k)}\phi_1^k(\theta)y^k} \equiv \sum_{j=0}^{m_1/d_1} \frac{a_{jd_1+k}\phi_1^{jd_1}(\theta)y^{jd_1}}{\pi^{v(a_k)}} \bmod (\pi).$$

For $\epsilon = \phi_1^{d_1}/\pi^{h_1}$ we have $v(\epsilon(\theta)) = v(\phi_1^{d_1}(\theta)/\pi^{h_1}) = 0$. Substitution of $\phi_1^{d_1}$ by $\epsilon\pi^{h_1}$ yields

$$\frac{\chi_1(\phi_1(\theta)y)}{\pi^{v(a_k)}\phi_1^k(\theta)y^k} \equiv \sum_{j=0}^{m_1/d_1} \frac{a_{jd_1+k}\pi^{jh_1}\epsilon^j y^{jd_1}}{\pi^{v(a_k)}} \bmod (\pi).$$

Replacing ϵy^{d_1} by z and considering the resulting polynomial over $\underline{K}$ yields the residual polynomial of S:

$$\underline{A}_1(z) := \sum_{j=0}^{m_1/d_1} \underline{a_{jd_1+k}\pi^{jh_1-v(a_k)}}z^j \in \underline{\mathsf{K}}[z].$$

For $\theta \in \Theta_1^*$ we have that $\underline{\phi_1^{d_1}(\theta)/\pi^{h_1}} \in \underline{\overline{\mathsf{K}}}$ is a zero of $\underline{A}_1$.

5.3 Unramified Extension I

Let f_1 be the least common multiple of the degrees of the irreducible factors of $\underline{A}$. The unramified extension K_1^* of K of degree f_1 is a subfield of L and $\underline{A}$ splits into linear factors over the residue class field $\underline{\mathsf{K}}_1^*$ of K_1.

Let Γ_1 be the set of zeros of $\underline{A}_1$ in K_1. By Proposition 1 we can obtain a proper factor of Φ over K_1 for each $\gamma \in \Gamma_1$. For some $\gamma_1 \in \Gamma_1$ let

$$\Theta_1 = \{\theta \in \Theta_1^* \mid \underline{\phi_1(\theta)^{d_1}/\pi^{h_1}} = \underline{\gamma}_1\}.$$

Our choice of γ_1 determines on which branch of the tree of partitions of Θ to follow.

5.4 Tamely Ramified Extension I

For all $\theta \in \Theta_1$ we have $v(\phi_1(\theta)) = \lambda_1 = h_1/d_1 = h_1/(e_1 p^{\mu_1^*})$ where $\gcd(h_1, d_1) = \gcd(e_1, p) = 1$.

If $e_1 = 1$ is a power of p, we set $\mathsf{K}_1 = \mathsf{K}_1^*$.

If $e_1 \neq 1$ is not a power of p, then the slope $-h_1/(e_1 p^{\mu_1^*})$ together with $\underline{\gamma}_1$ give enough information to provide a generating polynomial of a tamely ramified subfield $\mathsf{K}_1/\mathsf{K}_1^*$ of L. Let γ_1 be a lift $\underline{\gamma}_1$ to $\mathcal{O}_{\mathsf{K}_1}$. We have $\frac{\theta^{d_1}}{\pi^{h_1}} \sim \gamma_1$, so $\left(\theta^{p^{\mu_1}}\right)^{e_1} \sim \gamma_1\pi^{h_1}$. Therefore for $\delta = \theta^{p^{\mu_1}}$, we have $\delta^{e_1} \sim \gamma_1\pi^{h_1}$. The Newton polygon of $\tau^*(x) = x^{e_1} - \gamma_1\pi^{h_1}$ is a line of slope $-h_1/e_1$ with $\gcd(h_1, e_1) = 1$. By Proposition 2, if a and b are integers such that $ae_1 + bh_1 = 1$, the extension generated by τ^* is generated by the Eisenstein polynomial

$$\tau = x^{e_1} + (-1)^b(\gamma_1\pi^{h_1})^b\pi^{e_1 a} = x^{e_1} + (-1)^b\gamma_1^b\pi^{bh_1+ae_1} = x^{e_1} + (-\gamma_1)^b\pi.$$

As the splitting field of τ contains the e_1th roots of unity we continue the computations over the tamely ramified extension K_1 given by τ over the unramified extension of K_1^* that contains the e_1th roots of unity. Using Proposition 3 we can obtain the composite of the tamely ramified extensions of degree e_1 given by all $\gamma \in \Gamma_1$.

5.5 Next Approximation I

After the considerations in Sect. 5.4 above we can assume $\lambda_1 = h_1/d_1 = h_1/p^{\mu_1}$, as either d_1 was a power of p already or, if K_1 is a tamely ramified extension of K_1^*, recomputing the Newton polygon gives a segment with slope $-\lambda = -h_1/p^{\mu_1}$. In the latter case we also need to recompute the residual polynomial $\underline{A}$ in Sect. 5.2, which will be of higher degree. Again let Γ_1 be the set of zeros of $\underline{A}$, and $\underline{\gamma}_1 \in \Gamma_1$. Denote by γ_1 a lift of $\underline{\gamma}_1$ to $\mathscr{O}_{\mathsf{K}_1}$ and set

$$\Theta_1 = \{\theta \in \Theta_1^* \mid \underline{\phi_1^{d_1}(\theta)/\pi_{\mathsf{K}_1}^{h_1}} = \underline{\gamma}_1\}.$$

We have the relation $\phi_1^{p^{\mu_1}}(\theta) \sim \gamma_1 \pi_{\mathsf{K}_1}^{h_1}$ for all $\theta \in \Theta_1$. Now

$$\phi_2 = \phi_1^{p^{\mu_1}} - \gamma_1 \pi_{\mathsf{K}_1}^{h_1}$$

is an approximation to a polynomial generating the wildly ramified subextension of $\mathsf{K}_1(\theta)$ with

$$v_{\mathsf{K}_1}(\phi_2(\theta)) = v_{\mathsf{K}_1}(\phi_1^{p^{\mu_1}}(\theta) - \gamma_1 \pi_{\mathsf{K}_1}^{h_1}) > h_1 \geq h_1/d_1 = v_{\mathsf{K}_1}(\phi_1(\theta)).$$

5.6 Valuations I

Let $a = \sum_{j=0}^{d_1} a_j \phi_1^j \in \mathsf{K}_1[x]$ with $\deg a < \deg \phi_2 = d_1 = p^{\mu_1}$. As the valuations

$$v_{\mathsf{K}_1}(\phi_1(\theta)) = \frac{h_1}{d_1}, \ldots, v_{\mathsf{K}_1}(\phi_1^{d_1-1}(\theta)) = \frac{(d_1-1)h_1}{d_1}$$

are distinct (and not in $\mathbb{Z}$),

$$v_{\mathsf{K}_1}(a(\theta)) = \min_{0 \leq j \leq \deg d_1 - 1} v_{\mathsf{K}_1}(a_j(\theta)\phi_1^j(\theta)) = \min_{0 \leq j \leq \deg d_1 - 1} v_{\mathsf{K}_1}(a_j(\theta)) + j(h_1/d_1).$$

Furthermore, if we only consider the terms with valuation $v_{\mathsf{K}_1}(a(\theta))$ we obtain a polynomial that at θ is equivalent to $a(x)$. That is, if $v_{\mathsf{K}_1}(a_j(\theta)) + j(h_1/d_1) = v_{\mathsf{K}_1}(a(\theta))$, then we have $a(\theta) \sim a_j(\theta) + \phi_1^{j(h_1/d_1)}$ for $\theta \in \Theta_1$.

5.7 Polynomials with Given Valuations I

The data computed in the first iteration allows us, given $c \in \mathbb{Z}$ and $d \in \mathbb{N}$ with $\gcd(c, d) = 1$ and $d \mid d_1 = p^{\mu_1}$, to find $\psi(x) \in \mathsf{K}[x]$ with $v(\psi(\theta)) = \frac{c}{d}$ for $\theta \in \Theta_1$ and $\deg \psi < d_1$.

If $d = 1$, then $\psi(x) := \pi_{\mathsf{K}_1}^{s_\pi}$ with $s_{\pi_{\mathsf{K}_1}} = c$ has the property $v_{\mathsf{K}_1}(\psi(\theta)) = \frac{c}{d}$ for all $\theta \in \Theta_1$.

Otherwise d is a proper divisor of d_1. Find $s_1 \in \mathbb{Z}$ such that $s_1 h_1 \equiv \frac{c}{d} d_1 \bmod d_1$ and let $s_0 = \frac{c}{d} - v(\phi_1^{s_1}(\theta)) \in \mathbb{Z}$. Now for $\psi(x) := \pi_{\mathsf{K}_1}^{s_0} \phi_1(x)^{s_1} \in \mathsf{K}[x]$ we have $v(\psi(\theta)) = \frac{c}{d}$ for $\theta \in \Theta_1$.

5.8 Arithmetic I

We consider the arithmetic of polynomials of degree less than $d_1 = p^{\mu_1}$.

Let $a(x) = \sum_{i=0}^{d_1-1} a_i x^i$ and $\tau(x) = x^{s_1}\pi^{s_0}$ with $s_0, s_1 \in \mathbb{Z}$. Multiplication gives $a(x)\tau(x) = \sum_{i=0}^{d_1-1} a_i \pi_{\mathsf{K}_1}^{s_\pi} x^{i-s_1}$ which in general is a rational function or a polynomial of degree greater than $d_1 - 1$.

As $v(\phi_1^{d_1}(\theta)/\pi_{\mathsf{K}_1}^{h_1} - \gamma_1) = 0$ we have the relation

$$\phi_1^{d_1}(\theta) \sim \gamma_1 \pi_{\mathsf{K}_1}^{h_1}.$$

So by repeatedly substituting $\phi_1^{d_1}$ by $\gamma_1 \pi_{\mathsf{K}_1}^{h_1}$ we obtain a polynomial $b(x) \in \mathsf{K}[x]$ with $\deg b < d_1$ such that $b(\theta) \sim a(\theta)\psi(\theta)$.

6 The *u*th Iteration

We describe a general iteration of the algorithm. Let $t = (\phi_i, \lambda_i, \psi_i, y - \gamma_i)_{1 \le i \le u-1}$ be a wild extended type of Φ over K_{u-1} that is not complete. We write $\lambda_i = h_i / p^{\mu_i^*}$ with $\gcd(h_i, e_i p) = \gcd(e_i, p) = 1$ and set $M_{u-1} = \max\{\mu_i^* | 1 \le i \le u-1\}$ and $\mu_{u-1} = M_{u-1} - M_{u-2}$. Assume that we have found the next approximation $\phi_u \in \mathcal{O}_{\mathsf{K}_{u-1}}[x]$ to a generator of a wildly ramified extension with $\deg \phi_u = p^{M_{u-1}}$ and $v_{\mathsf{K}_u}(\phi_u(\theta)) > v_{\mathsf{K}_u}(\phi_{u-1}(\theta))$ for all $\theta \in \Theta_{u-1}$.

We assume that we have the following methods, which rely on the data computed in the previous steps. For each method the base case is described in Sect. 5 and the general case in this section. Because of the recursive nature of the algorithm we use forward references in our representation.

`Valuation` given $a(x) \in \mathsf{K}_u[x]$ with $\deg a < \deg \phi_u = p^{M_{u-1}}$ finds $v_{\mathsf{K}_u}(a(\theta))$ for $\theta \in \Theta_{u-1}$ (see Sects. 5.6, 6.7 and Algorithm 1).

`PolynomialWithValuation` given $c \in \mathbb{Z}$ and $d \in \mathbb{N}$ with $d \mid p^{M_u}$ finds $\psi(x) \in \mathsf{K}_u[x]$ with $\deg \psi < \deg \phi_u = p^{M_{u-1}}$ such that $v_{\mathsf{K}_u}(\psi(\theta)) = \frac{c}{d}$ for all $\theta \in \Theta_{u-1}$ (see Sects. 5.7, 6.8 and Algorithm 3).

Furthermore we assume that we have methods for arithmetic and reduction of polynomials of degree less than p^{M_u} in their representations as sums of power products (see Sects. 5.8, 6.9 and Algorithm 4 `reduce`).

In the uth iteration of the algorithm we investigate the properties of ϕ_u and construct the next approximation $\phi_{u+1} \in \mathscr{O}_{\mathsf{K}_u}[x]$ to a polynomial defining a wildly ramified subfield of the splitting field.

6.1 Newton Polygon II

We use the ϕ_u-expansion of Φ to find the valuations $v_{\mathsf{K}_{u-1}}(\phi_u(\theta))$ for $\theta \in \Theta_{u-1}$. Let $l_u = \lceil N/\deg \phi_u \rceil$ and $\Phi = \sum_{i=0}^{l_u} a_i \phi_u^i$ be the ϕ_u-expansion of Φ. For each root $\theta \in \Theta_{u-1}$ we have

$$\Phi(\theta) = \sum_{i=0}^{l_u} a_i(\theta)\phi_u^i(\theta) = 0.$$

Hence

$$\chi_u = \sum_{i=0}^{l_u} a_i(\theta) y^i \in \overline{\mathsf{K}}[y]$$

has the zeros $\phi_u(\theta)$ for $\theta \in \Theta_{u-1}$.

The method `Valuation` returns the valuations of the coefficients $a_i(\theta)$ of χ_u and with these the Newton polygon of χ_u yields the valuations of $\phi_u(\theta)$ for $\theta \in \Theta_{u-1}$. We obtain a partition of Θ_{u-1} into the subsets $\{\theta \in \Theta_{u-1} \mid v(\phi(\theta)) = \lambda\}$ where λ is the negative of the slope of a segment of the Newton polygon of χ_u. By Corollary 1 each segment of the Newton polygon of χ_u, and thus each set in the partition, corresponds to a factor of $\Phi(x)$.

Definition 7. The Newton polygon of χ_u is called the Newton polygon of Φ with respect to ϕ_u. It is also called a *Newton polygon of higher order* [8, 12].

6.2 Residual Polynomial II

Let S be a segment of the Newton Polygon of χ_u of length m_u with endpoints $(k, v_{\mathsf{K}_{u-1}}(a_k(\theta)))$ and $(k+m, v_{\mathsf{K}_{u-1}}(a_{k+m_u}(\theta)))$ for $\theta \in \Theta_{u-1}$. Let

$$\lambda_u = \frac{v_{\mathsf{K}_{u-1}}(a_k(\theta)) - v_{\mathsf{K}_{u-1}}(a_{k+n}(\theta))}{m_u} = \frac{h_u}{d_u},$$

where $\gcd(h_u, d_u) = 1$ and let $\Theta_u^* = \{\theta \in \Theta_{u-1} \mid v_{\mathsf{K}_{u-1}}(\phi_u(\theta)) = \lambda_u\}$. We have $|\Theta_u^*| = m_u \deg \phi_u$.

Let e_u and μ_u^* such that $d_u = e_u p^{\mu_u^*}$ with $\gcd(e_u, p) = 1$ and $M = \sum_{i=0}^{u-1} \mu_i = \max\{\mu_i^* \mid \mu_i^*\}$, $\nu = \min\{\mu^*, M\}$, and $\mu_u = \max\{\mu_u^* - M, 0\}$. The method `PolynomialWithValuation` gives $\psi_u \in \mathsf{K}_{u-1}[x]$ with

$$v_{\mathsf{K}_{u-1}}(\psi_u(\theta)) = v_{\mathsf{K}_{u-1}}\left(\phi_u^{e_u p^{\mu_u}}\right) = e_u p^{\mu_u} \lambda_u = h_u / p^{\nu}$$

for $\theta \in \Theta_u^*$. We have

$$\chi_u(\phi_u(\theta)) \sim \sum_{i=k}^{k+m} a_i(\theta) \phi_u^i(\theta) x^i \sim \sum_{j=0}^{m/(e_u p^{\mu_u})} a_{j e_u p^{\mu_u} + k}(\theta) \phi_u^{j e_u p^{\mu_u} + k}(\theta) x^{j e_u p^{\mu_u} + k}$$

The last equivalence holds, because the x-coordinates of the points on the segment of the Newton polygon are of the form $k + j e_u p^{\mu_u}$ $(0 \le j \le m/(e_u p^{\mu_u}))$. Division by $\phi_u^k y^k$ yields

$$\frac{\chi_u(\phi_u(\theta))}{\phi_u^k(\theta) y^k} \sim \sum_{j=0}^{m/(e_u p^{\mu_u})} a_{j e_u p^{\mu_u} + k}(\theta) \phi_u^{j e_u p^{\mu_u}}(\theta) y^{j e_u p^{\mu_u}}.$$

For $\gamma = \phi_u \theta^{e_u p^{\mu_u}} / \psi_u(\theta)$ we have $v_{\mathsf{K}_{u-1}}(\gamma) = v_{\mathsf{K}_{u-1}}(\phi_u^{e_u p^{\mu_u}}(\theta)/\psi_u(\theta)) = 0$. By substituting $\gamma \psi_u(\theta)$ for $\phi_u^{e_u p^{\mu_u}}(\theta)$ we get

$$\frac{\chi(\phi_u(\theta) y)}{\phi_u^k(\theta) y^k} \sim \sum_{j=0}^{m/(e_u p^{\mu_u})} a_{j e_u p^{\mu_u} + k}(\theta) (\gamma \psi_u^j(\theta)) y^{j e_u p^{\mu_u}}$$

The method `PolynomialWithValuation` gives a polynomial $\tau \in \mathsf{K}_{u-1}[x]$ with $v_{\mathsf{K}_{u-1}}(\tau(\theta)) = v_{\mathsf{K}_{u-1}}(a_k(\theta))$ for $\theta \in \Theta_{u-1}$. Replacing $\gamma y^{e_u p^{\mu_u}}$ by y and division by $\tau(\theta)$ yields

$$A(y) = \sum_{j=0}^{m/(e_u p^{\mu_u})} \frac{a_{j e_u p^{\mu_u} + k}(\theta) \psi_u^j(\theta)}{\tau(\theta)} y^j.$$

By construction $v_{\mathsf{K}_{u-1}}\left(\frac{a_{je_u p^{\mu_u}+k}(\theta)\psi_u^j(\theta)}{\tau(\theta)}\right) \geq 0$, in particular $v_{\mathsf{K}_{u-1}}\left(\frac{a_k(\theta)\psi_u(\theta)}{\tau(\theta)}\right) = 0$ and $v_{\mathsf{K}_{u-1}}\left(\frac{a_{k+m}(\theta)\psi_u^{m/(e_u p^{\mu_u})}(\theta)}{\tau(\theta)}\right) = 0$. So the polynomial $\underline{A}(z) \in \underline{\mathsf{K}}_{u-1}[z]$ has degree $m_u/(e_u p^{\mu_u})$. It is called the residual polynomial of S.

6.3 Unramified Extension II

Let f_u be the least common multiple of the degrees of the irreducible factors of $\underline{A}$ and let K_u^* be the unramified extension of K_{u-1} of degree f_u. Over $\underline{\mathsf{K}}_u^*$ the residual polynomial $\underline{A}$ splits into linear factors. We denote by Γ_u the set of lifts of zeros of $\underline{A}$ to K_u and partition Θ_u^* into the sets of the form

$$\Theta_u = \left\{\theta \in \Theta_u^* \;\middle|\; \underline{\frac{\phi_u^{e_u p^{\mu_u}}}{\psi_u}(\theta)} = \underline{\gamma}_u\right\}. \tag{4}$$

where $\gamma_u \in \Gamma_u$. We have $|\Theta_u| = g \deg \phi_u$, where g is the multiplicity of $\underline{\gamma}_u$ as a zero of $\underline{A}$.

6.4 Tamely Ramified Extension II

For $\theta \in \Theta_u$, we have $\underline{\left(\phi_u^{e_u p^{\mu_u}}/\psi_u\right)(\theta)} = \underline{\gamma_u}$, and therefore $\left(\phi_u^{p^{\mu^u}}(\theta)\right)^{e_u} \sim \gamma_u \psi_u(\theta)$. For $\tilde{\phi} = \phi^{p^{\mu_u}}$ we have $\tilde{\phi}^{e_u}(\theta) \sim \gamma_u \psi(\theta)$. As in Sect. 6.2 let $\nu = \min\{\mu_u^*, M_{u-1}\} = \mu_u^* - \mu_u$. Since

$$v_{\mathsf{K}_{u-1}}(\psi_u^{p^\nu}(\theta)) = p^\nu(e_u p^{\mu_u}\lambda_u) = p^{\mu_u^*-\mu_u} e_u p^{\mu_u} \frac{h_u}{e_u p^{\mu_u^*}} = h_u,$$

there is $\delta \in \mathscr{O}_{\mathsf{K}_{u-1}}^\times$ such that $\psi_u^{p^{\mu_u^*-\mu_u}}(\theta) \sim \delta \pi_{\mathsf{K}_{u-1}}^{h_u}$. With $\hat{\phi} = \tilde{\phi}^{p^\nu}$ we get

$$\hat{\phi}^{e_u}(\theta) = \tilde{\phi}^{p^\nu} \sim \gamma_u^{p^\nu} \psi^{p^{\nu_u}}(\theta) \sim \gamma_u^{p^\nu} \delta \pi^{h_u}.$$

Thus the roots of $\tau^*(x) = x^{e_u} - \gamma_u^{p^\nu} \delta \pi^{h_u}$ are in the splitting field L. Since the Newton polygon of τ^* is a line of slope $-h_u/e_u$ where $\gcd(h_u, e_u) = 1$, the polynomial τ^* defines a tamely ramified extension of K_u^* of degree e_u. By Proposition 2 it is generated by the Eisenstein polynomial

$$\tau(x) = x^{e_u} + (-1)^b(\gamma_u^{p^\nu}\delta\pi^{h_u})^b \pi^{e_u a} = x^{e_u} + (-1)^b \gamma_u^{bp^\nu} \delta^b \pi^{bh_u + ae_u} = x^{e_u} + (-\gamma_u^{p^\nu}\delta)^b \pi$$

where a and b are integers such that $ae_u + bh_u = 1$.

6.5 Wildly Ramified Extension

Now either d_u was a power of p already or, after extending K_u^* the slope of the segment of the Newton polygon of χ_u corresponding to S now is $-h_u/p^{\mu_u^*}$ over K_u. If K_u is a tamely ramified extension of K_u^*, we would need to recompute the associated polynomial $\underline{A}$ and to obtain a residual polynomial of higher degree. Hence we assume $\lambda_u = h_u/d_u = h_u/p^{\mu_u^*}$.

If $|\Theta_u| = \deg \phi_u = 1$, we have reached a leaf of the tree of partitions.

If $|\Theta_u| > \deg \phi_u$, we continue with constructing a next approximation to a polynomial that generates the wildly ramified part of $\mathsf{K}_u(\theta)$ for $\theta \in \Theta_u$.

If $|\Theta_u| = \deg \phi_u = p^{M_{u-1}} \neq 1$, then ϕ_u is an approximation to a unique irreducible factor $\hat{\phi}$ of degree $p^{M_{u-1}}$ of Φ over K_u. We obtain $\hat{\phi}$ using single factor lifting [9], which generates a totally and wildly ramified extension M of K_u over which Φ has at least one linear factor. Now for $1 \leq i \leq u-1$ with $\mu_i > 0$ the data in $t = (\phi_i, \lambda_i, \psi_i, y - \gamma_i)_{1 \leq i \leq u-1}$ and the slopes $-\lambda_i$, and thus ψ_i and $\underline{A}_i$ are not correct over M. Thus we continue our computations with the type $t = (\phi_i, \lambda_i, \psi_i, y - \gamma_i)_{1 \leq i \leq j}$ over M, where $1 \leq j \leq u-1$ is such that $\mu_i = 0$ for $1 \leq i \leq j$.

6.6 The Next Approximation II

As above in Sect. 6.5 we assume that $\lambda_u = h_u/d_u = h_u/p^{\mu_u^*}$ and as in Sect. 6.2 let $\psi_u \in \mathsf{K}_{u-1}[x]$ with $v(\psi_u(\theta)) = h_u/p^{\mu_u^*}$ for $\theta \in \Theta_u \subseteq \Theta_{u-1}$. If Γ_u denotes the set of zeros of the residual polynomial and $\gamma_u \in \Gamma_u$, then $\phi_u^{p^{\mu_u^*}}(\theta) \sim \gamma_u \psi_u$ for all $\theta \in \Theta_u$. The polynomial

$$\phi_{u+1} = \phi_u^{p^{\mu_u^*}} - \gamma_u \psi_u$$

is an approximation to a polynomial that generates the wildly ramified subextension of $\mathsf{K}_u(\theta)$ with

$$v_{\mathsf{K}_u}(\phi_u(\theta)) = v_{\mathsf{K}_u}(\phi_u^{p^{\mu_u^*}}(\theta) - \gamma_u \psi_u) = v_{\mathsf{K}_u}(\psi_u) = h_u/p^{\nu} \geq h_u/d_u = v_{\mathsf{K}_u}(\phi_u(\theta))$$

for all $\theta \in \Theta_u$.

6.7 Valuations II

For $b(x) \in \mathsf{K}_{u-1}[x]$ with $\deg b < p^{M_{u-1}}$ the method `Valuation` yields $v_{\mathsf{K}_{u-1}}(a(\theta))$ for $\theta \in \Theta_u \subset \Theta_{u-1}$. Let $a \in \mathsf{K}_u[x]$ with $\deg a < p^{M_u}$ and

$m = \lceil \deg a / \deg \phi_u \rceil$. Let $a = \sum_{j=0}^{m} a_j \phi_u^j$ with $\deg a_j < \deg \phi_u = p^{M_{u-1}}$ be the ϕ_u-expansion of a. As the valuations

$$v_{\mathsf{K}_u}(\phi_u(\theta)) = \frac{h_1}{d_u}, \dots, v_{\mathsf{K}_u}(\phi_u^{p^{\mu_u}-1}(\theta)) = \frac{(p^{\mu_u}-1)h_u}{d_u}$$

are distinct (and not in $\frac{1}{p^{M_{u-1}}}\mathbb{Z}$) we have

$$v_{\mathsf{K}_u}(a(\theta)) = \min_{0 \le j \le m} v_{\mathsf{K}_u}\left(a_j(\theta)\phi_u^j(\theta)\right) = \min_{0 \le j \le m} v_{\mathsf{K}_u}\left(a_j(\theta) + j(h_u/p^{\mu_u})\right).$$

Furthermore, if we only consider the terms with valuation $v_{\mathsf{K}_u}(a(\theta))$, we obtain a polynomial that at θ is equivalent to $a(x)$. That is, for $J = \{j \mid v_{\mathsf{K}_u}(a_j) + jh_u/d_u = v_{\mathsf{K}_u}(a(\theta))\}$ and $b(x) = \sum_{j \in J} a_j(x)\phi_u^j(x)$ we have $a(\theta) \sim b(\theta)$ for $\theta \in \Theta_u$.

6.8 Polynomials with Given Valuations II

Let $c \in \mathbb{Z}$ and $d \in \mathbb{N}$ with $d \le M_u$. We describe how $\psi(x) \in \mathsf{K}_u[x]$ with $v_{\mathsf{K}_u}(\psi(\theta)) = \frac{c}{p^d}$ and $\deg \psi < \deg \phi_u = p^{M_{u-1}}$ can be constructed. Assume that for $c' \in \mathbb{Z}$ and $d' \in \mathbb{N}$ with $d' < M_{u-1}$ we can find $\psi'(x) \in \mathsf{K}[x]$ with $v_{\mathsf{K}_u}(\psi'(\theta)) = \frac{c'}{p^{d'}}$ for $\theta \in \Theta_u \subseteq \Theta_{u-1}$.

If $d < M_{u-1}$, then we can find $\psi(x)$ by our assumption. Otherwise we have $M_{u-1} < d \le M_u$ and we find $s_u \in \mathbb{Z}$, $0 \le s_u < p^{\mu_u}$ such that

$$s_u h_u \equiv c p^{M_u - d} \bmod p^{\mu_u}$$

and set $\frac{c'}{p^{d'}} = \frac{c}{p^d} - s_u v_{\mathsf{K}_u}(\phi_u)$ in lowest terms. As $d' < M_{u-1}$ the assumption yields $\psi'(x) \in \mathsf{K}_u[x]$ with $v_{\mathsf{K}_u}(\psi'(\theta)) = \frac{c'}{p^{d'}}$. Thus we get $\psi(x) = \phi_u^{s_u}(x)\psi'(x)$ with $v_{\mathsf{K}_u}(\psi(\theta)) = \frac{c}{p^d}$ and $\deg \psi < p^{M_u}$.

6.9 Arithmetic II

We consider the arithmetic of polynomials of degree less than p^{M_u}. Clearly addition and subtraction of two such polynomials again yield polynomials of degree less than p^{M_u}. We assume that methods for handling polynomials of degree less than $p^{M_{u-1}}$ are available. That is, given $a(x) \in \mathsf{K}_{u-1}$ and $b(x) \in \mathsf{K}_{u-1}$ we can find a polynomial $c \in \mathsf{K}_{u-1}[x]$ with $\deg c < p^{M_{u-1}}$ such that $c(\theta) \sim a(\theta)b(\theta)$ for $\theta \in \Theta_u \subseteq \Theta_{u-1}$.

Let $a(x) = \sum_{i=0}^{p^{M_u}-1} a_i(x)\phi_u^i$ and $b = \phi_u^s b'$ with $s \in \mathbb{Z}$, $b' \in \mathsf{K}_u[x]$ of degree less than $p^{M_{u-1}}$. Multiplication gives $a(x)b(x) = \sum_{i=0}^{p^{M_u}-1} a_i b' \phi_u^{i+s}$ which in general is a rational function or a polynomial of degree greater than $p^{M_u}-1$. We have $\underline{\phi_u^{e_u}(\theta)/\psi_u(\theta)} = \underline{\gamma}_u$, thus

$$\phi_u^{e_u}(\theta) \sim \gamma\psi_u(\theta)$$

for any $\gamma \in \mathsf{K}_u$ with $\underline{\gamma} = \underline{\gamma}_u$. Repeated substitution of $\phi_u^{p^{\mu_u}}$ by $\gamma\psi_u$ reduces the exponents of ϕ_u to s' with $0 \le s' < p^{\mu_u}$. The coefficient of $\phi_u^{s'}$ now is the product of polynomials of degree less than p^{M_u}, which can be reduced to a polynomial of degree less than p^{M_u} by our assumption. Thus we obtain a polynomial $b(x) \in \mathsf{K}[x]$ with $\deg b < p^{M_u}$ such that $b(\theta) \sim a(\theta)\psi_u(\theta)$. If $v_{\mathsf{K}_u}(a(\theta)) = 0$ recursive application of these reductions yield $\beta \in \mathsf{K}_u$ with $a(\theta) \sim \beta$.

7 Algorithms

In our formulation of the algorithm we add a fifth component to the extended types from Definition 4, making all the information from previous iterations of the algorithm needed in later iterations readily available. We denote the subfield of the splitting field of Φ over which we are working at all times by L. So in this section types are of the form

$$t = (\phi_i, \lambda_i, \psi_i, \gamma_i, \mu_i)_{1\le i\le u} \tag{5}$$

with $\phi_i \in \mathscr{O}_{\mathsf{L}}[x]$, $\lambda_i = \frac{h}{e_i p^{\mu_i^*}} \in \mathbb{Q}$ with $\gcd(h, e_i p) = \gcd(e_i, p) = 1$, $\gamma_i \in \underline{\mathsf{L}}$, $\mu_i \in \mathbb{N}$ such that $\mu_i = \max\{\mu_i^* - M_i, 0\}$ where $M_i = \max\{\mu_j^* \mid 1 \le j \le i-1\} = \sum_{j=1}^{i-1} \mu_j$. By $\Theta = \Theta_u$ we denote the set of zeros that corresponds to t as in (4).

In an implementation of the algorithm, the methods described below operate on representations of polynomials as nested ϕ_i-expansions ($1 \le i \le u$). To avoid having to write down these somewhat involved data structures, we use polynomials to formulate the input and the output of the methods. Sections 5.8 and 6.9 yield these methods:

`div`(t, a, b) Given $a \in \mathsf{K}[x]$ of degree less than p^{M_u} and $b(x) = \phi_u^{s_u}\dots\phi_1^{s_1}\pi^{s_\pi}$ where $s_i < e_i$ we find $c \in \mathsf{K}[x]$ with $\deg c < \deg\phi_u$ such that $a(\theta)/b(\theta) \sim c(\theta)$ for all $\theta \in \Theta_u$

`mult`(t, a, b) Given $a, b \in \mathsf{K}[x]$ of degree less than p^{M_u} we find $c \in \mathsf{K}[x]$ with $\deg c < \deg\phi_u$ such that $a(\theta)b(\theta) \sim c(\theta)$ for all $\theta \in \Theta_u$.

`pow`(t, a, n) Given $a \in \mathsf{K}[x]$ of degree less than p^{M_u} we find $c(x) \in \mathsf{K}[x]$ with $\deg c < \deg\phi_u$ such that $a^n(\theta) \sim c(\theta)$ for all $\theta \in \Theta_u$.

Furthermore we write `divmod` for the function that for $a, b \in \mathbb{Z}$ returns the quotient and remainder of the division of a by b.

We first give auxiliary algorithms for the computation of $v_t(a) = v(a(\theta))$ for $\theta \in \Theta_u$, the Newton polygon of Φ with respect to ϕ, polynomials with given valuations,

Algorithm 1 `Valuation`

```
Input:  A local field $\mathsf{L}$, type $(\phi_i, \lambda_i, \psi_i, \gamma_i, \mu_i)_{1\le i\le u}$ over $\mathsf{L}$, and $a(x) \in \mathsf{L}[x]$.
Output: Valuation $v_t(a)$.

• If $a \in \mathsf{L}$: Return $v_{\mathsf{L}}(a)$.
• Find the $\phi_{u-1}$-expansion of $a(x) = \sum_{j=0}^{\lceil \deg a/\deg \phi_u \rceil} a_j(x)\phi_u^j(x)$.
• Return min $\left\{\texttt{Valuation}\left(\mathsf{L}, (\phi_j, \lambda_j, \psi_j, \gamma_j, \mu_j)_{1\le j\le u-1}, a_j\right) + j\lambda_{u-1} \;\middle|\; 1 \le i \le \lceil \frac{\deg a}{\deg \phi_{u-1}} \rceil\right\}$
```

Algorithm 2 `NewtonPolygonSegments`

```
Input:  A local field $\mathsf{L}$, $\Phi \in \mathsf{L}[x]$, a type $t = (\phi_i, \lambda_i, \psi_i, \gamma_i, \mu_i)_{1\le i\le u}$ over $\mathsf{L}$,
        and $\phi \in \mathscr{O}_{\mathsf{L}}[x]$
Output: Set of Segments $S$ of the Newton polygon of $\Phi$ with respect to $\phi$.

• Find the $\phi$-expansion $\Phi = \sum_{i=0}^{m} a_i\phi^i$ where $m = \lceil \deg \Phi / \deg \phi \rceil$.
• Find $v_i = \texttt{Valuation}(\mathsf{L}, t, a_i)$ for $0 \le i \le m$.
• Construct the lower convex hull of the set of points $\{(i, v_i) \mid 1 \le i \le m\}$.
• Return the set $S$ of segments of this broken line.
```

the reduction of elements represented as power products of polynomials, and the computation of residues and residual polynomials. This is followed by the algorithm for the splitting field.

We use Algorithm 1 `Valuation` to compute $v_{\mathsf{L}}(a(\theta))$ for $\theta \in \Theta_u$. It follows from the discussions in Sects. 5.6 and 6.7 that to find $v_{\mathsf{L}}(a(\theta))$ for $\theta \in \Theta_u$ we only need the type $t = (\phi_i, \lambda_i, \rho_i)_{1\le i\le u}$ and not θ. We thus obtain one of the valuations of polynomial rings as classified by MacLane in [10]. We write $v_t(a)$ for the valuation computed by the algorithm and have $v_t(a) = v_{KK_u}(a(\theta))$.

Algorithm 2 `NewtonPolygonSegments` returns the set of segments of the Newton polygon of Φ with respect to ϕ as described in Sects. 5.1 and 6.1.

Given a type t as in (5) and $w \in \frac{1}{p^{M_u}}\mathbb{Z}$, Algorithm 3 `PolynomialWithValuation` returns a polynomial ψ such that $v_t(\psi) = w$ as described in Sects. 5.7 and 6.8. See [18], [17, Algorithm 14] or [9, Section 4] for a general version of this algorithm.

In Sects. 5.8 and 6.9 we have described how a product $\prod_{i=1}^{u} \phi_i^{s_i}(x)$ can be reduced such that $s_i < p^{\mu_i}$ for $1 \le i \le u$. Algorithm 4 `reduce` conducts this reduction recursively. Because, for $1 \le i \le u$ the valuations of $\phi_i^{s_i}$ with $s_i < p^{\mu_i}$ are linearly independent, there is only one reduced representation of each class of some $a \in \mathsf{L}[x]$ with respect to the equivalence relation from Definition 1. Thus if $v_t(a) = 0$, then `reduce`$(a) \in \mathsf{L}$.

The residual polynomial of a segment of a Newton polygon of higher order is computed in Algorithm 5 `ResidualPolynomial`.

Algorithm 6 `SplittingField` computes the splitting field of a polynomial. If $\mu_u = 0$ and $\deg \phi = 1$ and the multiplicity of the zero γ of $\underline{A}$ is one, then ϕ

Algorithm 3 `PolynomialWithValuation`

```
Input:  A type $(\phi_i, \lambda_i, \psi_i, \gamma_i, \mu_i)_{1 \le i \le u}$ and $\frac{c}{p^d} \in \mathbb{Q}$ with $d \le \sum_{i=1}^{u-1} \mu_i$.
Output: $\psi(x) \in \mathsf{K}[x]$ with $\deg \psi < \deg \phi_u$ and $v_{\mathsf{L}}(\psi(\theta)) = \frac{c}{p^d}$.

• If $d = 0$: Return $\pi^c$.
• $M \leftarrow \sum_{i=1}^{u-2} \mu_i$.
• If $d \le M$: Return PolynomialWithValuation $\left((\phi_i, \lambda_i, \psi_i, \gamma_i, \mu_i)_{1 \le i \le u-1}, \frac{c}{p^d}\right)$.
• Find $0 \le s < p^{\mu_{u-1}}$ such that $sh_{u-1} \equiv cp^{M+\mu_{u-1}-d} \bmod p^{\mu_{u-1}}$.
• Return $\phi_u^s(x) \cdot$ PolynomialWithValuation $\left((\phi_i, \lambda_i, \psi_i, \gamma_i, \mu_i)_{1 \le i \le u-1}, \frac{c}{p^d} - s\lambda_{u-1}\right)$.
```

Algorithm 4 `reduce`

```
Input:  A type $(\phi_i, \lambda_i, \psi_i, \gamma_i, \mu_i)_{1 \le i \le u}$ and $a(x) = \phi_u^{r_u} \cdot \prod_{i=1}^{u-1} \phi_i^{r_i} \cdot \delta \in \mathsf{L}[x]$
        with $\delta \in \mathsf{K}$.
Output: $b(x) = \phi_u^{s_u} c(x) \in \mathsf{L}[x]$ with $\deg c < \deg \phi_u$, $0 \le s_u < p^{\mu_u}$, and
        $a(\theta) \sim b(\theta)$ for $\theta \in \Theta_u$.

• If $a \in \mathsf{L}$: Return $a$.
• $s, d \leftarrow$ divmod$(r_u, p^{\mu_u})$
• Return $\phi_u^s \cdot \gamma_u^d \cdot \psi_u^d \cdot$ reduce $\left((\phi_i, \lambda_i, \psi_i, \gamma_i, \mu_i)_{1 \le i \le u-1}, \prod_{i=1}^{u-1} \phi_i^{r_i} \cdot \delta\right)$.
```

Algorithm 5 `ResidualPolynomial`

```
Input:  A type $(\phi_i, \lambda_i, \psi_i, \gamma_i, \mu_i)_{1 \le i \le u}$, a segment $S$ of the Newton polygon of
        $\Phi$ with respect to $\phi$, and $\psi$ with $v_t(\psi) = \lambda_u = ep^{\min\{\mu, M_u\}}$ where
        $-h/(ep^\mu)$ is the slope of $S$.
Output: The residual polynomial $\underline{A}$ of $S$.

• Let $\Phi = \sum_{i=0}^{\lceil N/\deg \phi_u \rceil} a_i \phi^i$ be the $\phi$-expansion of $\Phi(x)$.
• Let $m$ be the length of $S$.
• $\tau \leftarrow$ PolynomialWithValuation$(t, \nu)$ where $\nu$ is the $y$-coordinate of the first point of $S$.
• $\underline{A}(z) \leftarrow \sum_{j=0}^{m/e_u}$ reduce$(t,$ mult$(t, a_{k+e \cdot i}(x),$ div$(t,$ pow$(t, \psi(x), j), \tau(x))))y^j$.
• return $\underline{A}$.
```

Algorithm 6 `SplittingField`

```
        Input:   $\Phi \in \mathcal{O}_{\mathsf{K}}[x]$ monic and square-free
        Output:  Splitting field of $\Phi$

• Initialize $\mathsf{L} \leftarrow \mathsf{K}$ and $T \leftarrow \{(x, \cdot, \cdot, \cdot, \cdot)\}$
• While $T$ is non-empty:

   1. Choose a type $t = (\phi_i, \lambda_i, \psi_i, \gamma_i, \mu_i)_{1 \le i \le u}$ from $T$.
   2. Remove $t$ from $T$.
   3. $M \leftarrow \sum_{i=1}^{u-1} \mu_i$
   4. For $S \in$ NewtonPolygonSegments $(\Phi(x), t, \phi_u(x))$:

      a. Let $\lambda_u = \frac{h}{ep^{\mu^*}}$ with $\gcd(h, pe) = \gcd(e, p) = 1$ be the negative of the slope of $S$.
      b. $\mu_u \leftarrow \max\{\mu^* - M, 0\}$ and $\nu \leftarrow \min\{\mu^*, M\}$
      c. $\psi_u \leftarrow$ PolynomialWithValuation$(t, h/p^\nu)$
      d. $\underline{A} \leftarrow$ ResidualPolynomial $(t, S, \psi)$
      e. Find $f_0$ minimal with $e \mid (|\underline{\mathsf{L}}^{f_0}| - 1)$.
      f. If $f = \operatorname{lcm}\{f_0, \deg \underline{\rho} \mid \underline{\rho}$ irreducible factor of $\underline{A}\} > 1$:
         continue over unramified extension

            · Replace $\mathsf{L}$ by the unramified extension of $\mathsf{L}$ of degree $f$.

      g. If the length of $S$ is one and $\deg \phi_u > 1$:
         continue over wildly ramified extension

            · Let $\hat{\phi}$ be a lift of $\phi_u$ to a factor of $\Phi$.
            · Replace $\mathsf{L}$ by $\mathsf{L}[x]/(\hat{\phi})$.
            · Insert $t$ into $T$.
            · Replace each $(\phi_j, \lambda_j, \psi_j, \gamma_j, \mu_j)_j \in T$ by $(\phi_1, \cdot, \cdot, \cdot, \cdot)$
            · Exit for loop.

      h. If $e > 1$:
         continue over tamely ramified extension

            · Find $\delta \in \mathcal{O}_{\mathsf{K}}$ such that $\psi^{p^\nu} = \delta\pi^h$.
            · Find $a, b \in \mathbb{Z}$ be such that $ae + bh = 1$.
            · Replace $\mathsf{L}$ by the composite of $\mathsf{L}[x]/(x^e + (-\gamma^{p^\nu}\delta)^b \pi_L)$ where the $\gamma$ are lifts
              of the roots of $\underline{A}$ in $\underline{\mathsf{L}}$.
            · Insert $t$ into $T$.
            · Exit for loop.

      i. For all roots $\underline{\gamma}$ of $\underline{A}$ in $\underline{\mathsf{L}}$:

            · Let $\gamma$ be a lift of $\underline{\gamma}$ to $\mathcal{O}_{\mathsf{L}}$.
            · If $\mu_u > 0$:
              more wild ramification found, a Montes step

                  Insert $t$ with $(\phi_u^{\mu_u} - \gamma\psi_u, \cdot, \cdot, \cdot, \cdot)$ appended into $T$.

            · Else if $\deg \phi > 1$ or the multiplicity of $\gamma$ is greater than 1:
              valuation of $\phi$ increases, but not its degree, an improvement step

                  Insert $t$ with its last member replaced by $(\phi_u - \gamma\psi_u, \cdot, \cdot, \cdot, \cdot)$ into $T$.

• Return $L$.
```

is an approximation to a linear factor and therefore discarded. In a type we denote non-assigned components by $\cdot$. In each iteration of the algorithm we start with a type whose last member has only the first component set and then fill in the other components. At the end of each iteration the last member of all types again only have the first component assigned. We start with $t = (x, \cdot, \cdot, \cdot, \cdot)$. The types in the algorithm are at all times optimal. In step **4i.** when the current iteration only yields a ϕ with a higher valuation at $\theta \in \Theta$ we replace the last component of the type t by the current $(\phi, \cdot, \cdot, \cdot, \cdot)$; this is called an improvement step. If the degree of ϕ in increases we append $(\phi, \cdot, \cdot, \cdot, \cdot)$ to t; this is called a Montes step.

Remark 1. For better readability of the algorithm we have excluded some obvious improvements. When we continue over a tamely ramified extension in **5h.**, instead of exiting the for loop and recomputing the Newton polygon in **4.** we can adjust the slopes of the segments of the Newton polygon, which over the tamely ramified extension of degree e do not have e in the denominator anymore and continue in **4c**. When choosing t from T in **1.** a speedup may be achieved by first processing the longest type as this avoids discarding information about wildly ramified extensions in **4h**. If $\phi \in \mathcal{O}_{\mathsf{K}}[x]$ is Eisenstein, the maximal tamely ramified subfield of the splitting field of ϕ can be obtained from [6, Theorem 9.1].

7.1 Termination

The termination of the algorithm is assured by the following theorem.

Theorem 2 ([16, Proposition 4.1]). *Let $\Phi(x) \in \mathcal{O}_{\mathsf{K}}[x]$ be square-free and let Θ_0 be the set of zeros of $\Phi(x)$ in $\overline{\mathsf{K}}$. Let $\phi(x) \in \mathsf{K}[x]$ such that the degree of any irreducible factor of $\Phi(x)$ is greater than or equal to* $\deg \phi$. *If* $(\deg \Phi) \cdot v(\phi(\theta_j)) > 2v(\operatorname{disc} \Phi)$ *for all $\theta \in \Theta_0$, then* $\deg \phi = \deg \Phi$ *and $\Phi(x)$ is irreducible over* K.

By Theorem 2 the polynomial $\Phi(x)$ is irreducible if we find a monic $\phi(x) \in \mathcal{O}_{\mathsf{K}}[x]$ such that $N v(\phi_u(\theta)) > 2v(\operatorname{disc} \Phi)$ for some $u \in \mathbb{N}$. In every iteration of the algorithm the increase from $v(\phi_u)$ to $v(\phi_{u+1})$ is at least $1/N$. Thus the algorithm terminates after at most $v(\operatorname{disc} \Phi)$ iterations.

7.2 Representation of Extensions

Extensions of local fields are often represented as a tower of an totally ramified extension over an unramified extension. Although some computer algebra systems (for example, Magma [3]) allow the construction of arbitrary towers of extensions, in practice it is more efficient to work over smaller towers. We represent our extensions as a totally and wildly ramified extension over a totally and tamely ramified extension over an unramified extension and insert new extensions into the corresponding subfield in the tower.

Ramified extensions are usually given by Eisenstein polynomials, and indeed for the tamely ramified extensions in our algorithm we explicitly give these. For the wildly ramified extensions, lifting the approximation to an irreducible factor ϕ by single factor lifting yields a generating polynomial $\hat{\phi}$ that is not Eisenstein in general. Algorithm 3 can be used to compute $\Pi \in \mathsf{L}[x]$ with $v_t(\Pi) = 1/p^M$ where $p^M = \deg \phi$. The characteristic polynomial (see Definition 2) of $\Pi \in \mathsf{L}[x]/(\hat{\phi})$ is the desired Eisenstein polynomial. It can be computed either using linear algebra methods as in [5] or, in the p-adic case, using Newton relations.

References

1. Awtrey C, Miles N, Milstead J, Shill C, Strosnider E (to appear) Galois groups of degree 14 2-adic fields, Involve. http://msp.org/scripts/coming.php?jpath=involve
2. Bauch J, Nart E, Stainsby D (2013) Complexity of OM factorizations. LMS J Comput Math 16:139–171
3. Cannon JJ et al (2014) The computer algebra system Magma. University of Sydney. http://magma.maths.usyd.edu.au/magma/
4. Ford D (1978) On the computation of the maximal order in a Dedekind domain. Ph.D. dissertation, Ohio State University
5. Ford D, Pauli S, Roblot X-F (2002) A fast algorithm for polynomial factorization over $\mathbb{Q}_p$. J Théor Nombres Bordeaux 14(1):151–169
6. Greve C, Pauli S (2012) Galois groups of Eisenstein polynomials whose ramification polygon has one side. Int J Number Theory 8(6):1401–1424
7. Guàrdia J, Montes J, Nart E (2010) Okutsu invariants and Newton polygons. Acta Arithmet 145:83–108
8. Guàrdia J, Montes J, Nart E (2012) Newton polygons of higher order in algebraic number theory. Trans Am Math Soc 354(1):361–416
9. Guardia J, Nart E, Pauli S (2012) Single-factor lifting and factorization of polynomials over local fields. J Symb Comput 47(11):1318–1346
10. MacLane S (1936) A construction for absolute values in polynomial rings. Trans Am Math Soc 40:363–395
11. MacLane S (1936) A construction for prime ideals as absolute values of an algebraic field. Duke Math J 2:493–510
12. Montes J (1999) Polígonos de Newton de orden superior y aplicaciones aritméticas. Ph.D. thesis, Universitat de Barcelona
13. Montes J, Nart E (1992) On a theorem of Ore. J Algebra 146:318–334
14. Okutsu K (1982) Construction of integral basis, I, II, III, and IV. Proc Jpn Acad Ser A 58:47–49, 87–89, 117–119, and 167–169
15. Ore Ö (1928) Newtonsche Polygone in der Theorie der algebraischen Körper. Math Ann 99:84–117
16. Pauli S (2001) Factoring polynomials over local fields. J Symb Comput 32:533–547
17. Pauli S (2010) Factoring polynomials over local fields II. In: Hanrot G, Morain F, Thomé E (eds) Algorithmic number theory, 9th international symposium, ANTS-IX, Nancy, France, July 19–23, 2010. Lecture notes computer science, vol 6197. Springer, Berlin, pp 301–315
18. Pauli S, Sinclair B (2014) A guide to OM algorithms (in preparation)
19. Zassenhaus H (1969) On Hensel factorization I. J Number Theory 1:291–311

Weight Loss Through Bariatric Surgery: Some Issues

Robert Stoesen, Kristin McLamb, Laurie Deaton, and Sat Gupta

1 Introduction

In this study, we are presented with data for 122 patients who underwent laparoscopic gastric banding (Lap-Band) surgery for treatment of clinically severe obesity at a Cone Health System facility between June 2008 and May 2009. Questions of interest include whether the amount of weight lost following the procedure is correlated with the number of support group sessions attended by patients and with various demographic data for the patients. Because of the retrospective nature of this study, we were unable to make a baseline adjustment at this point.

We define obesity in a medical context, discuss the medical treatment of obesity, and in particular examine the nature of bariatric surgical procedures and the history of their development.

For the Cone Health System data, we conducted statistical analyses of the data using a variety of techniques. We were not able to determine a significant relationship between the number of support group meetings attended by patients and the amount of weight they lost, nor did we find that age, sex, or race were contributing factors in such a relationship. We also did not find that the mean amount of weight lost was significantly different for African Americans compared to Caucasians, nor did we find that the mean amount of weight lost was significantly different for females compared to males. We found that Caucasian-identified patients may

R. Stoesen • S. Gupta (✉)
Department of Mathematics and Statistics, UNC Greensboro, Greensboro, NC, USA
e-mail: r_stoese@uncg.edu; sngupta@uncg.edu

K. McLamb • L. Deaton
Cone Health, Greensboro, NC, USA
e-mail: kristin.mclamb@conehealth.com; laurie.deaton@conehealth.com

J. Rychtář et al. (eds.), *Collaborative Mathematics and Statistics Research*, Springer Proceedings in Mathematics & Statistics 109, DOI 10.1007/978-3-319-11125-4_12

be more likely to attend support group meetings than African American-identified patients. We did not find that one gender was more likely to attend support group meetings than another.

2 Background: Obesity and Its Treatment

In recent decades, obesity has become a topic of increasing concern both in public discourse and in the medical and public health communities. The popular media is replete with references to conditions of overweight and obesity among the populace. In the present context, we define obesity medically as a chronic disease. Specifically, obesity is a condition of excess body fat, defined in terms of body mass index (BMI). BMI is calculated as weight in kilograms divided by height in meters squared, rounded to one decimal place. For adults of typical height range in the USA, obesity is defined by the Centers for Disease Control as BMI greater than or equal to 30. Similar definitions of obesity are used by medical authorities around the globe.

As a chronic disease, obesity is associated with multiple factors, including environmental (social and cultural), genetic, physiologic, metabolic, behavioral, and psychological componentsİ [2].

Obesity is associated with high rates of mortality and morbidity. Mortality refers to the rate of deaths from a disease, while morbidity refers to the rate of incidence of a disease. The World Health Organization notes that overweight and obesity are the fifth leading risk for global deaths [11].

The prevalence of obesity is high in the USA. The National Center for Health Statistics of the Centers for Disease Control and Prevention has reported that in 2009–2010, 35.7 % of USA adults were obese, a total of over 78 million US adults. Additionally, 16.9 % of US children and adolescents were obese, a total of about 12.5 million US young people. It was also reported that over the prior decade, the prevalence of obesity had increased among men and boys, but not among women and girls [10]. Beyond the USA, the World Health Organization reports that worldwide obesity has nearly doubled since 1980 to a total, in 2008, of upwards of 1.4 billion adults (defined as age 20 or older) worldwide, comprising over 200 million men and 300 million women [11].

The disease of obesity is associated with numerous comorbidities, that is, other diseases that are observed accompanying a condition. A partial list of comorbidities associated with obesity include type II diabetes, most forms of cancer, all cardiovascular diseases, asthma, gallbladder disease, osteoarthritis, and chronic back pain. Researchers have concluded that the comorbidities associated with obesity results in a disease "likely (to) contribute substantially to the burden of chronic health conditions." [9].

The phenomena surrounding obesity continues to be studied. Recently, public health researchers suggested that the rate of mortality associated with obesity in the USA might be as much as four times higher than previously believed. If true, this raises concerns about a reversal in the trend of lengthening life spans in the USA.

Other experts dispute this finding [6]. Nevertheless, a review of the literature and reports from organizations confirms a widely accepted medical and public health consensus that obesity is a serious and deadly condition that merits the serious attention of the clinical and research communities of the medical and public health fields, as well as society at large.

There is also ongoing research about obesity on a cellular level. Research has revealed that fat cells expand adipose tissue by enlargement and by division. When these cells reach their maximum size, and intake continues to exceed expenditure (caloric burn), the fat cells will increase in quantity. When fat is lost, the cells do shrink, but the number of them remains the same. This poses great difficulty in reducing the total amount of excess adipose tissue. Additionally, more than 600 genes, markers, and chromosomes have been linked with obesity causing conditions/diseases such as hypertension, diabetes, lipid abnormalities, and body fat distribution [4].

According to the Mayo Clinic, a variety of strategies are used in the medical treatment of obesity, including changes in diet, programs of exercise and activity, behavioral change, prescription weight loss medications, and weight-loss surgery. These treatments are to be distinguished from weight loss schemes such as fad diets and non-approved medications [8]. Weight loss surgery has become increasingly prevalent and accessible in the USA. Overall, the success of surgery as a treatment option has led to an increased number of surgeries. In 2004, approximately 144,000 surgeries were performed compared to over 225,000 currently. Insurance companies are also recognizing the benefit of providing coverage for such surgical procedures. With the average cost of surgery ranging between $12,000 and $35,000, this is far below the costs associated with ongoing care and treatment of obesity related diseases such as diabetes and hypertension [7].

Many members of the public are familiar with celebrity figures who have had weight loss surgery. NBC Today Show host and weather forecaster Al Roker has spoken publicly about his weight loss surgery and subsequent dramatic weight loss on television and in a published memoir of his career. Other celebrities and public figures have also had high profile weight loss following surgery, including New Jersey Governor Chris Christie, actress Roseanne Barr, and singer Carnie Wilson, among others.

Weight loss surgery, known as bariatric surgery, is surgery intended to restrict a patient's caloric intake or nutrient absorption to effect weight loss over time. Bariatric surgical procedures are divided into three categories of procedure: restrictive, malabsorptive, and combination. Restrictive procedures reduce the patient's caloric intake by decreasing the ability to consume food beyond a certain quantity. Malabsorptive surgery involves creating a bypass in some portion of the patient's nutrient absorption circuit, thereby decreasing the absorption of calories from food already consumed. Combination procedures involve both strategies [3].

Experimental bariatric surgery was first performed in the period following World War II, and techniques were developed and modified in the ensuing decades. The earliest bariatric surgery procedures were malabsorptive techniques developed in the early 1950s, but with limited success. In the 1960s and 1970s, new malabsorptive procedures were developed, but patients developed significant complications. Combination procedures in the 1970s had fewer complications, but patients developed significant long-term side effects and protein malnutrition [3].

Laparoscopic, or minimally invasive, surgical techniques for bariatric procedures were developed in the early 1990s. In 1992, the first placement of a nonadjustable gastric band in a patient was performed. As surgeons became more experienced with laparoscopic procedures and parity was achieved with open procedures with respect to successful results, the laparoscopic surgeries grew to exceed the number of open procedures such that laparoscopic techniques are described as the greatest contributor to the increase in weight loss surgeries of the previous decade [3].

Laparoscopic banding or Lap-Band surgery (a trademark owned by Allergan, Inc.) involves surgical placement of an inflatable silicone band around the upper part of the stomach. An attached port extends from the band to a terminal end outside the body below the abdomen through which a doctor or medical professional can inject a saline solution to inflate the band, thereby constricting the stomach. With this restrictive surgical measure in place, a sensation of fullness after eating occurs more quickly in the patient, leading to less food intake and hence weight loss over time. Patients are told they may expect to lose up to one pound per week in the first year following surgery [1].

The surgery is approved for adult patients only. The implanted device is intended to remain in the body over the course of the patient's life, in order to maintain desired weight after weight loss and to prevent a return to previous eating habits. However, it can be removed if the patient elects to do so.

Unlike some other bariatric procedures, Lap-Band surgery requires patients to return to a doctor's office on a regular basis for monitoring and adjustment of the device. For instance, a patient may require that the amount of saline solution in the device be increased or decreased for their comfort or to make the device more effective [5].

3 Research Question

Researchers in the nursing department at the Cone Health organization in Greensboro, NC, were interested in the issue of social support for patients following Lap-Band bariatric surgery. In particular, their interest was in whether a positive relationship could be ascertained for patients attendance in support groups in the period following surgery and their success in achieving weight loss.

The Cone Health researchers presented data for patients who had undergone Lap-Band surgery at their facilities between June 2008 and May 2009. The patients were counseled prior to their surgery about the importance of social support during the course of their post-surgery weight loss. Patients were offered the opportunity to attend free, optional bariatric support group meetings following their surgery. The support group is led by a dedicated registered nurse specializing in care of bariatric patients. The meetings included invited speakers to address the groups on topics of interest, such as nutrition or healthy lifestyle choices, and also included the opportunity for patients to discuss matters of concern to them in an informal setting.

The patients' weights and BMI were recorded approximately 2 weeks prior to surgery and again approximately 1 year following surgery. Some patients attended support group meetings during the 1-year period following their surgery, and their attendance was recorded.

In considering this data, we examined the change in pre- and post-surgery weight measurements and considered various questions, taking into account the number of support group meetings attended and demographic variables of age, race, and gender.

4 Data Analysis and Results

Data was provided for 122 patients. Of those patients, 119 had lost weight 1 year after having the surgery, which is consistent with published studies indicating that gastric banding is highly effective in achieving immediate weight loss for virtually all patients [5].

During the year following surgery, patients attended between 0 and 15 support group meetings. An initial examination of the data reveals that 73 % of these patients did not attend a support group meeting. Of the remaining patients, 11 % attended 2 or 3 meetings, 6 % attended 4 or 5 meetings, and 3% attended 7–15 meetings (see Fig. 1). We then compared the amount of weight lost by those in the sample who attended support group meetings with the amount lost by those who did not.

Figure 1 shows the percentage of the 121 patients under consideration for the number of meetings attended.

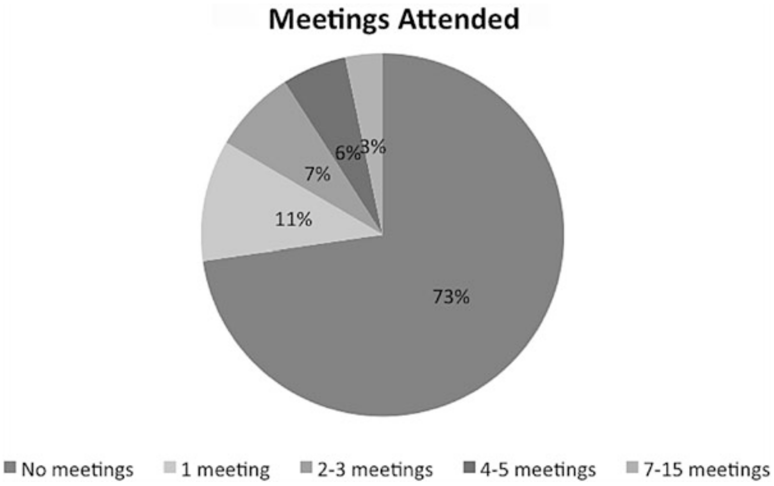


Fig. 1 Percentage of the patients attending meetings by number of meetings attended

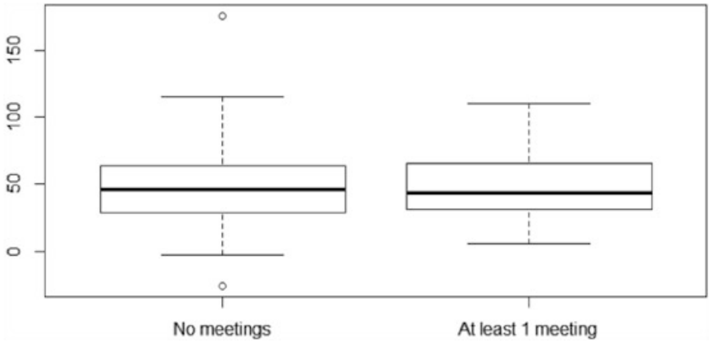


Fig. 2 Comparison of patient weight loss in pounds by meeting attendance (no meetings versus at least one)

The study was strictly retrospective. We did not have other information that might have shed additional light on why most patients did not attend support group meetings. For instance, we did not have information about these patients geographic distribution, modes of transportation, employment and family status, or access to other social support networks.

The following analysis relied upon the data that was provided, which was limited to race, gender, and change in weight and BMI after 1 year. Other information, such as the name of the patient's surgeon and the specific dates of their surgeries, was not used.

The box plots in Fig. 2 represent the distribution of weight lost in pounds by patients who attended no support group meetings and those who attended at least one meeting. The y-axis represents the amount of weight lost after 1 year following surgery.

The median figures for the two groups, 46.30 for those attending no meetings and 44.20 for those attending at least one appear quite similar, as do the surrounding quartiles. The spread of the distribution for the non-attending group is larger, however, so we consider only those patients who attended at least one support group in the following analysis.

After transforming the data and checking for normality, a simple linear regression was performed for explanatory variable number of support groups attended and the log of the response variable change in weight (weight loss). However, the number of support groups attended was not found to be significant, with a p-value of 0.362. One patient was observed to be an outlier with a weight loss of 175.4 pounds. Removing this outlier and refitting the regression model produced a similarly large p-value of 0.304. We also attempted a step-wise multiple regression incorporating the additional variable of age and the factors of race and sex, but the additional components were not found to be significant and were discarded by the algorithm.

The patients were then divided into categories of having attended 0, 1, 2–3, and 4 or greater support group sessions, and tests were performed with regard to change in weight. The normal Q–Q plot of sample quantiles versus theoretical quantiles

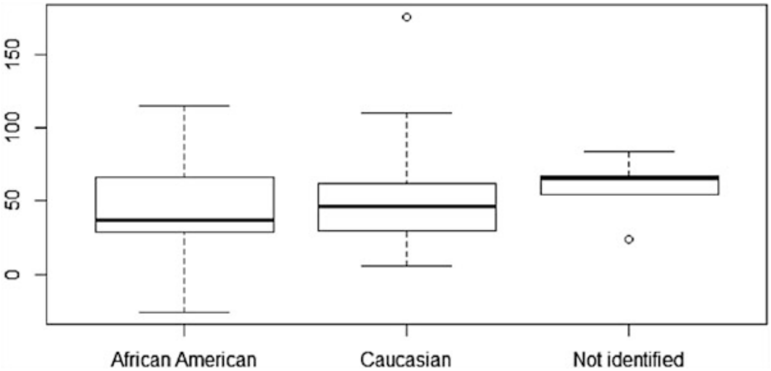


Fig. 3 Comparison of patient weight loss in pounds by race

suggested that despite some slight skewness, the assumption of normality was not violated. Bartlett's Test for homogeneity of variance yielded a p-value of 0.5726, strong evidence for not rejecting the null hypothesis of equal variances for these categories. Analysis of variance (ANOVA) yielded an F-statistic of 0.3422 and does not suggest any significant difference among categories with regard to weight loss. Dividing patients into binary categories of having attended or not attended support groups, another linear regression was performed. However, again, this was not found to be significant with regard to weight lost, yielding a p-value of 0.543.

Next we considered demographic factors. The box plots in Fig. 3 represent the distribution of weight lost in pounds among patients by race.

We observed that the mean weight loss for Caucasian-identified patients is 49.32302 pounds, while the mean weight loss for African American-identified patients is 45.17333, a difference of 4.14969 pounds. We used a t-test to determine if there is a significant difference in these means. However, the result indicates there is no difference in the means, with a p-value of 0.4932 and a 95 % confidence interval of $(-16.232753, 7.933373)$.

Next we compared patients identified as Caucasian and African American with patients not identified by race, after first removing the outliers observed in Fig. 3. For Caucasians compared to non-identified patients, the t-test yields a p-value of 0.03886 and a 95 % confidence interval of $(-37.713198, -1.607273)$, moderate evidence of some difference in means. For African Americans compared to non-identified patients, the t-test yields a p-value of 0.02223 and a 95 % confidence interval of $(-40.610170, -4.043164)$, slightly stronger evidence of a difference in means. However, we note that the non-identified group comprised only five patients, four when the outlier was removed.

The box plots in Fig. 4 represent the distribution of weight lost among patients by sex.

We observed that mean weight loss for female patients is 47.62076 pounds, while the mean weight loss for male patients is 55.67500, a difference of 8.05424 pounds. We used a t-test to determine if there is a significant difference these means.

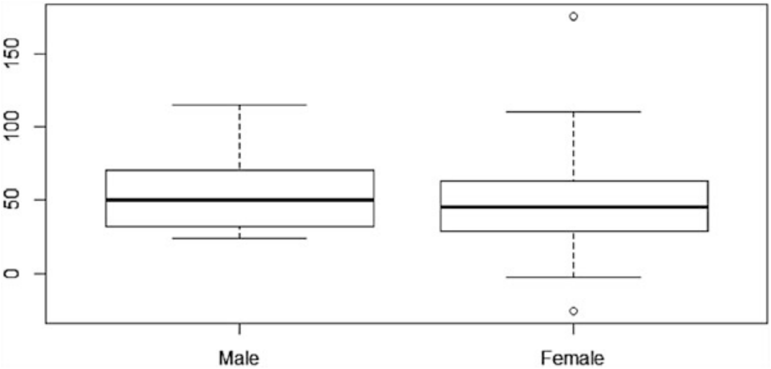


Fig. 4 Comparison of patient weight loss in pounds by gender

However, the result indicates there is no difference in the means, with a *p*-value of 0.2976 and a 95 % confidence interval of (–23.608675, 7.500289).

Having determined that the data does not suggest a significant relationship between the amount of weight lost and the number of support groups attended, race, sex, or age, we next considered how the number of support groups attended correlated with some of the demographic variables. Given the small number of observations for some categories, Pearson's chi-square test was not feasible for this analysis. We computed the Wilcoxon rank-sum test for the data. For race (not considering the three patients whose race was not identified), the two-sided *p*-value of 0.07059 and 95 % confidence interval of $(-8.58922\times10^{-6}, 5.974492\times10^{-6})$ may indicate some significant difference in the two populations (Caucasian and African American), with Caucasian-identified patients attending a larger number of support group meetings. For sex, the *p*-value of 0.8452 and 95 % confidence interval of $(-5.707147 \times 10^{-5}, 2.41973 \times 10^{-5})$ indicate no significant difference in the two populations (female and male).

Conclusions

For this sample, we are not able to determine a significant relationship between the number of support group meetings attended by patients and the amount of weight they lost, nor did we find that age, sex, or race were contributing factors in such a relationship. We also did not find that the mean amount of weight lost was significantly different for African Americans compared to Caucasians, nor did we find that the mean amount of weight lost was significantly different for females compared to males. We found that Caucasian-identified patients may be more likely to attend support group

(continued)

meetings than African American-identified patients. We did not find that one gender was more likely to attend support group meetings than another.

In examining patient centered care, health care providers must research what options will provide the most success and satisfaction in the journey to weight loss. Live, face-to-face meetings may not be the most optimal form of support and education for many patients. In the quest for weight loss success, the findings from this study can be used to encourage future research on methods that can best be employed to aid patients following Lap-Band bariatric surgery.

References

1. Allergan, Inc. (2014) The Lap-Band system. www.lapband.com. Cited 7 March 2014
2. Bagchi D, Preuss H (2012) Obesity: epidemiology, pathophysiology, and prevention, 2nd edn. CRC Press, New York
3. Baker M (2011) The history and evolution of bariatric surgery procedures. Surg Clin North Am 91:1181–1201
4. Dalton S (2006) Obesity trends: past, present and future. Topics Clin Nutr 21(2):76–94
5. Fielding G, Ren C (2005) Laparoscopic adjustable gastric band. Surg Clin North Am 85:129–140
6. Healy M (2013) Obesity's death toll could be higher than believed, Study Says. Los Angeles Times. 15 August 2013
7. Ide P, Fitzgerald-O'Shea C, Lautz D (2013) Implementing a bariatric surgery program. AORN J 97(2):195–209. doi:10.1016/j.aorn.2012.11.018
8. Mayo Clinic Staff (2014) Obesity. http://www.mayoclinic.org/diseases-conditions/obesity/basics/definition/con-20014834. Cited 10 March 2014
9. Must A, Spadano J, Coakley E, Field A, Colditz G, Dietz W (1999) The disease burden associated with overweight and obesity. JAMA 282(16):1523–1529. doi:10.1001/jama.282.16.1523
10. Ogden CL, Carroll MD, Flegal KM (2012) Prevalence of obesity in the United States, 2009–2010. Natl Center Health Stat Data Brief 82:1–8
11. World Health Organization (2013) Obesity and overweight. Fact Sheet 311

On the Stability of Solutions to a Phase Transition Model

Heather Hardeman and Stephen Robinson

1 Introduction

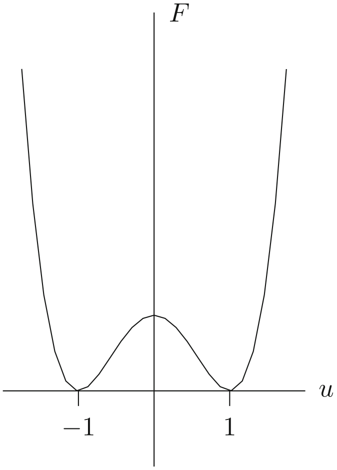
In this paper, we study the functional

$$J_\epsilon(u) := \frac{\epsilon^2}{2}\int_0^1 |u_x|^2 dx + \int_0^1 F(u)dx, \quad u \in W^{1,2}(0,1)$$

under suitable assumptions on $F : \mathbb{R} \longrightarrow \mathbb{R}$. This functional represents the total free energy of a phase transition model. In nature, energy seeks a minimum. As such, we can see that the first integral wants the slope of u to be flat whereas the second integral is pushing towards the minimum values of F. With this in mind, we want to consider a function F which has the following shape:

H. Hardeman • S. Robinson (✉)
Department of Mathematics, Wake Forest University, Winston Salem, NC 27109, USA
e-mail: hardhk12@wfu.edu; sbr@wfu.edu

J. Rychtář et al. (eds.), *Collaborative Mathematics and Statistics Research*, Springer
Proceedings in Mathematics & Statistics 109, DOI 10.1007/978-3-319-11125-4_13



A function with this shape captures the idea of an object with two different states of being. A good example of phase transition is water phasing from a liquid to a solid.

For our purposes, two different cases of F which have this shape are the classical case $F(u) = (1 - u^2)^2$ and the non-classical case $F(u) = (1 - u^2)^\alpha$ where $1 < \alpha < 2$. While the classical and non-classical cases appear indistinguishable from a modeling perspective, they have very different consequences. In fact, the non-classical case has features which the classical case lacks. These features of the non-classical case may help explain observed phenomena such as "slow dynamics." As such, our focus will be on the non-classical case of F.

In [1], Drábek, Manásevich, and Tákăc discovered all of the critical points of the functional J_ϵ. These critical points of the functional correspond to solutions of the Neumann boundary value problem

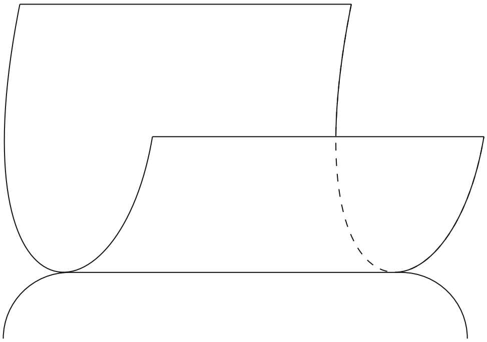
$$\begin{aligned} -\epsilon^2 u_{xx} + F'(u) = &\quad 0, \\ u_x(0) = u_x(1) = &\quad 0. \end{aligned} \tag{1}$$

These solutions are also all the stationary solutions of the bistable equation

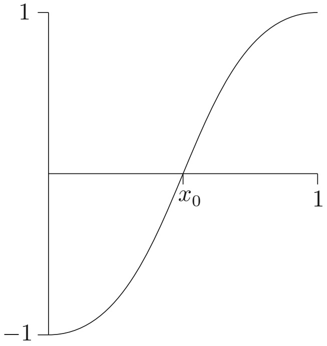
$$\begin{aligned} u_t - \epsilon^2 u_{xx} + F'(u) = &\quad 0, \\ u_x(0,t) = u_x(1,t) = &\quad 0, \end{aligned} \tag{2}$$

with the initial condition $u(x,0) = u_0(x)$.

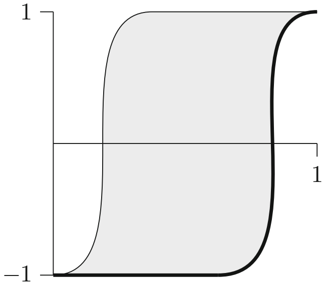
Note that ± 1 are global minima of J_ϵ. It is not hard to see that 0 is a saddle point. Since the critical point 0 is a saddle point, we also know that it is an unstable solution of (2). In [1], the authors also describe manifolds of solutions corresponding to single node solutions, two-node solutions, etc. For example, the set of single node solutions can be visualized as a trough. In [2], Drábek and Robinson verified that the interior of each of these manifolds consists of local minima.



The image above depicts a "graph" of J_ϵ with critical points lying along the base of the trough. These critical points are the one-node solutions of (1). They are solutions which cross the x-axis only once at some point x_0. The solution is equal to 1 on $[1 - \delta_1, 1]$ and it is equal to -1 on $[0, \delta_2]$ for $\delta_1, \delta_2 > 0$. The figure below gives an example of a solution located along the bottom of the trough.



By moving the point x_0 where the solution crosses the x-axis, we obtain these other solutions which lie along the base of the trough. For each x_0, there is a unique solution by Theorem 7.1 in [1]. The image below depicts two different cases. One is a solution located at the left side of the trough and the other is a solution from the right side of the trough. The rest of the solutions which are in the base reside in the shaded region between these two solutions.



Given that the local geometry of $J_\epsilon(u)$ takes the shape of a trough for the one-node solutions, the question of stability arises. As such, the purpose of this paper is to study the stability of these solutions. In particular, we want to know what set of initial conditions for (2) will be drawn to the bottom of the trough. Another way to say this is that we are studying the stable manifold for the trough.

In this paper, we will focus specifically on the case when the solution is symmetric via $u(1-x) = -u(x)$. This occurs when $x_0 = \frac{1}{2}$. Our consideration of the symmetric case results in the following theorem.

Theorem 1 (Hardeman, Robinson (2013)). *Let u be the one-node solution of the time independent problem* (1) *satisfying* $u(\frac{1}{2}) = 0$ *and* $u(0) = -1$. *Assume* $-1 < u_0(x) \le \beta u(x)$ *for* $x \in [0, 1/2]$ *and some* $0 < \beta < 1 - \frac{1}{n}$. *Assume* $u_0(x) = -u_0$ $(1-x)$. *Then, the solution,* $h(x,t)$, *of* (2) *satisfies* $\lim_{t \longrightarrow \infty} h(x,t) = u(x)$ *(uniformly).*

Theorem 1 shows that the symmetric initial condition is in the stable manifold, i.e., the solution corresponding to the symmetric initial condition converges to a point in the trough.

We will utilize the method of upper and lower solutions to prove Theorem 1. A common assumption for this method is that the forcing function satisfies a Lipschitz condition, but that is not the case for our problem. Given the nature of the non-classical F, we will need to deal with the unboundedness of the slope of the function F' near -1 and 1 which we will do by truncating problem (2).

2 Background

First, we will consider each of these definitions and theorems in terms of the following generalized problems:

$$\begin{aligned} \epsilon^2 u_{xx} + f(u) = 0 \quad & \text{in } (a,b) \\ \alpha \frac{\partial u}{\partial \nu} + \beta u = 0 \quad & \text{at } x = a, b \end{aligned} \tag{3}$$

where f is Lipschitz continuous and $\alpha, \beta \ge 0$ such that $\alpha + \beta > 0$. Also, we want to consider

$$\begin{aligned} u_t - \epsilon^2 u_{xx} - f(u) &= 0 \text{ in } (a,b) \text{ and } t > 0 \\ u(x,0) &= u_0(x) \\ \alpha \frac{\partial u}{\partial \nu} + \beta u &= 0 \quad \text{at } x = a, b \end{aligned} \tag{4}$$

with f, α, and β defined as above.

Before we prove the main theorem, we must first consider some definitions. These are standard results for upper and lower solutions. We mention them for the sake of completeness.

Definition. A smooth function $\overline{u}$ is an upper solution of the boundary value problem (4) if

$$\begin{aligned} &\overline{u}_t - \epsilon^2 \overline{u}_{xx} \geq f(\overline{u}) \text{ in } (a,b) \\ &\alpha \frac{\partial \overline{u}}{\partial \nu} + \beta \overline{u} \geq 0 \text{ on the lateral boundaries, and} \\ &\quad \overline{u}(x,0) \geq u_0(x). \end{aligned}$$

with the same conditions for f, α, and β as for (3).

Note that by reversing the inequalities, we obtain the definition of a lower solution. Now, we will consider some theorems.

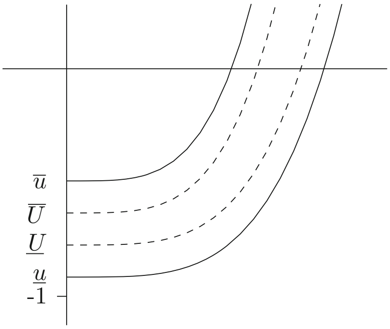
Theorem 2. *If there exists $\underline{u}$, $\overline{u}$ as defined above such that $\underline{u} \leq \overline{u}$, then the boundary value problem* (3) *has a solution u such that $\underline{u} \leq u \leq \overline{u}$.*

Lemma 2.1. *Let $\overline{u}(x), \underline{u}(x)$ be upper and lower solutions of* (3) *and let $\overline{U}(x,t)$, $\underline{U}(x,t)$ be solutions of* (4) *corresponding to the initial conditions $u_0 = \overline{u}$ and $u_0 = \underline{u}$, respectively. Assume f is a C^1-function in $\langle \underline{u}, \overline{u} \rangle$. Let $\underline{u} \leq \overline{u}$. Then, for $x \in [a,b]$, $\overline{U}(x,t)$ is non-increasing in t, $\underline{U}(x,t)$ is non-decreasing in t, and*

$$\underline{u}(x) \leq \underline{U}(x,t) \leq \overline{U}(x,t) \leq \overline{u}(x) \quad \text{in } \mathcal{D}.$$

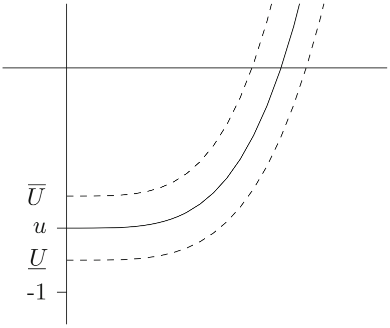
If $\underline{u}$ (resp. $\overline{u}$) is not a solution of (3)*, then $\underline{U}$ is strictly increasing (resp. $\overline{U}$ is strictly decreasing) in t.*

Note that $\langle \underline{u}, \overline{u} \rangle$ is the interval of functions bounded below by $\underline{u}$ and bounded above by $\overline{u}$.



With the above picture in mind and [3], Lemma 2.1 tells us that if we have an upper and a lower solution with a solution $U(x,t)$ between them, then depending on the initial condition of $U(x,t)$, $U(x,t)$ is non-increasing or non-decreasing in time. For two solutions $\overline{U}(x,t)$ and $\underline{U}(x,t)$ between $\overline{u}$ and $\underline{u}$ with $\overline{u}$ the initial condition of $\overline{U}(x,t)$ and $\underline{u}$ the initial condition of $\underline{U}(x,t)$, then Lemma 2.1 also tells us $\overline{u}(x) \geq \overline{U}(x,t) \geq \underline{U}(x,t) \geq \underline{u}(x)$. This result will be a key factor in proving Theorem 1.

Lemma 2.2. *Let the hypotheses of the previous lemma hold, and let $u(x,t)$ be the solution of* (4) *with $u_0 \in \langle \underline{u}, \overline{u} \rangle$. Then, $\underline{U}(x,t) \leq u(x,t) \leq \overline{U}(x,t)$ in $\mathcal{D}$.*



Now, Lemma 2.2 tells us that given solutions $\overline{U}$ and $\underline{U}$ of (4), then, for a solution $u(x,t)$ of (4) with $u_0 \in \langle \underline{u}, \overline{u} \rangle$, $u(x,t)$ lies between $\overline{U}$ and $\underline{U}$. We will see later how this lemma in conjunction with Lemma 2.1 helps us obtain the proof for Theorem 1.

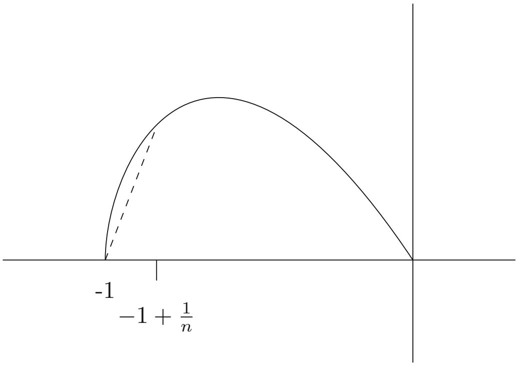
3 Theorem

Recall that the non-classical $F(u) = (1-u^2)^\alpha$ for $1 < \alpha < 2$. Therefore, as u approaches ±1, the slope of F' becomes unbounded. It follows that F' is not Lipschitz. To avoid the unboundedness of F' at these points, we will truncate F'.

So we define

$$f_n(u) = \begin{cases} F'(u) & -1 + \frac{1}{n} \leq u \leq 0 \\ nF'(-1 + \frac{1}{n})(u + 1) & -1 < u < -1 + \frac{1}{n}. \end{cases}$$

Consider the following graph of $f_n(u)$ on the interval $(-1, 0)$.



The dotted line represents the point when $f_n(u)$ no longer equals F'. Also, $f_n'(u)$ is bounded as it approaches $u = -1$. However, it follows that as $n \longrightarrow \infty$, $f_n(u)$ converges to $F'(u)$ uniformly. Observe that in the following proofs, we use only the fact that $f_n \leq F'$ and $f_n \longrightarrow F'$ uniformly as $n \longrightarrow \infty$. Therefore we could replace f_n by a smoother approximation.

First, we note that if the initial condition is symmetric around $x_0 = \frac{1}{2}$, then so is the solution of (2). This allows us to reduce our argument to the following problem. Further, we truncate F to get

$$\begin{aligned} u_t &= \epsilon^2 u_{xx} - f_n(u) \\ u_x(0, t) &= 0 = u\left(\tfrac{1}{2}, t\right) \\ u(x, 0) &= u_0(x). \end{aligned} \tag{5}$$

Note that when $u \in [-1 + \frac{1}{n}, 0]$, (5) reduces to (4).

Now, in order to use Lemmas 2.1 and 2.2, we must find an upper and a lower solution for (5). Since we are working with Neumann boundary conditions, we will prove a function is an upper solution using the following inequalities:

$$\begin{aligned} \overline{u}(\tfrac{1}{2}) &\geq 0 \\ \overline{u}_x(0) &\leq 0 \\ \epsilon^2 \overline{u}_{xx} &\leq 0. \end{aligned} \tag{6}$$

For the lower solutions, we only need to reverse the inequalities. Note that we have just chosen values for α and β from the upper solution definition in order to get (6).

Lemma 3.1. *Let $u(x)$ be the symmetric one-node solution to* (1) *satisfying $u(0) = -1$ and $u(\frac{1}{2}) = 0$. Then, $\overline{u} = \beta u(x)$ is an upper solution for the problem* (5) *with $0 < \beta < 1 - \frac{1}{n}$.*

Before we prove this lemma, we want to recall that $u(x)$ is unique by Theorem 7.1 from [1].

Proof. Let $0 < \beta < 1 - \frac{1}{n}$. Consider $\beta u(x)$. We must show that $\beta u(x)$ satisfies the following conditions.

First, notice $\beta u(\frac{1}{2}) = 0$. Therefore, $\overline{u}(\frac{1}{2}) \geq 0$. Now, observe $(\beta u(x))_x = \beta u_x(x)$. Recall $u_x(0) = 0$ since $u(x)$ is a solution of (1). Thus, $\overline{u}_x(0) \leq 0$. Finally, note that $(\beta u(x))_{xx} = \beta u_{xx}(x)$, so we have $\epsilon^2(\beta u_{xx}(x)) = \beta(\epsilon^2 u_{xx}(x)) = \beta F'(u(x))$. Since $0 < \beta < 1 - \frac{1}{n}$, then

$$\begin{aligned}
(u(x))^2 &\geq \beta^2(u(x))^2,\\
\text{so } -(u(x))^2 &\leq -\beta^2(u(x))^2.\\
\text{Also, } 1-(u(x))^2 &\leq 1-\beta^2(u(x))^2,\\
\text{and } (1-(u(x))^2)^{\alpha-1} &\leq (1-\beta^2(u(x))^2)^{\alpha-1} \quad \text{since } 1<\alpha<2;\\
\text{thus, } -2\alpha\beta u(x)(1-(u(x))^2)^{\alpha-1} &\leq -2\alpha\beta u(x)(1-\beta^2(u(x))^2)^{\alpha-1}\\
&\qquad \text{since } u(x) \leq 0 \text{ on } \left[0, \tfrac{1}{2}\right].\\
\text{Then } \beta F'(u(x)) &\leq F'(\beta u(x)).
\end{aligned}$$

Therefore, $\epsilon^2 \beta u_{xx}(x) \leq f_n(\beta u_{xx}(x))$ since $f_n(u) = F'(u)$ for $u \in (-1 + \frac{1}{n}, 0)$. Hence, $\beta u(x)$ is an upper solution for (5) when $0 < \beta < 1 - \frac{1}{n}$. □

Lemma 3.2. *Define u as in Lemma 3.1. Then $\underline{u} = u(x)$ is a lower solution of the problem* (5).

Proof. Let $\epsilon > 0$ be given. First, we notice $u(\frac{1}{2}) = 0$, so $\underline{u}(\frac{1}{2}) = 0$. Also, recall $u_x(0) = 0$ as a solution of (1). Therefore, $\underline{u}_x(0) = 0$. Finally, note that

$$0 = \epsilon^2 u_{xx}(x) - F'(u(x)).$$

since $u(x)$ is a solution of (1). By a calculus argument, we have that F' is concave down on the interval $(-1, -1+\frac{1}{n})$. Hence, $F'(u(x)) \geq f_n(u(x))$ for all x. Therefore,

$$\begin{aligned}
0 &= \epsilon^2 u_{xx}(x) - F'(u(x))\\
&\leq \epsilon^2 u_{xx}(x) - f_n(u(x)).
\end{aligned}$$

Then, $\epsilon^2 u_{xx}(x) \geq f_n(u(x))$, so $\epsilon^2 \underline{u}_{xx} \geq f_n(\underline{u})$. Therefore, $u(x)$ is a lower solution of (5). □

We want to also note that the interval $[\underline{u}, \overline{u}] = [u(x), \beta u(x)]$ contains no solutions to (1) other than $\underline{u} = u(x)$ [1]. With this in mind, we are now prepared to prove Theorem 1 for the symmetric one-node case.

Proof. Let $h_n(x,t)$ be the solution of (5). Let $\overline{U}_n(x,t)$ and $\underline{U}_n(x,t)$ be solutions of (5) with initial conditions $u_0(x) = \overline{u}(x)$ and $u_0(x) = \underline{u}(x)$, respectively. By Lemmas 3.1 and 3.2, $\beta u(x)$ is an upper solution of (5) for $0 < \beta < 1 - \frac{1}{n}$ and $u(x)$ is a lower solution of (5), respectively. By Theorem 2, we know that such a solution $h_n(x,t)$ exists and is bounded between $\underline{u}$ and $\overline{u}$. Also, note that $u(x) \le \beta u(x)$ since $u(x) \le 0$ for $x \in [0, \frac{1}{2}]$. Thus, by Lemma 2.2,

$$\underline{U}_n(x,t) \le h_n(x,t) \le \overline{U}_n(x,t)$$

for all x and t. Notice $\underline{U}_n(x,t) \ge u(x)$. So, we have

$$u(x) \le h_n(x,t) \le \overline{U}_n(x,t).$$

Observe that for any given n and all (x,t) such that $\overline{U}_n(x,t) \ge -\frac{1}{n} + 1$ we have $f_n(\overline{U}_n(x,t)) \equiv F'(\overline{U}_n(x,t))$. By uniqueness, we have $\overline{U}(x,t) \equiv \overline{U}_n(x,t)$ for such (x,t). Similarly, $h_n(x,t) = h(x,t)$ for such (x,t). Therefore,

$$u(x) \le h(x,t) \le \overline{U}(x,t).$$

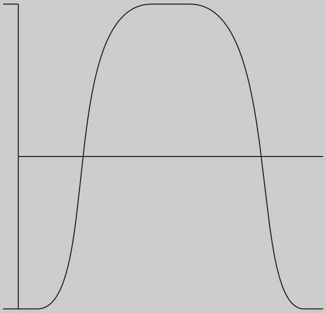
Now, by Lemma 2.1, we know $\overline{U}(x,t)$ is non-increasing in t. Also, $\overline{U}(x,t)$ is bounded below by the solution $u(x)$. By [1], we know that $\overline{U}$ converges uniformly to a solution of (1) as $t \longrightarrow \infty$, and we know that $u(x)$ is the only solution in $\{v : u \le v \le \overline{u}\}$. Therefore, $\lim_{t \longrightarrow \infty} \overline{U}(x,t) = u(x)$ uniformly. Hence, by the squeeze theorem, $\lim_{t \longrightarrow \infty} h(x,t) = u(x)$ uniformly. □

Conclusion
In conclusion, we shall discuss some open questions on which we are working that relate to our previous work.

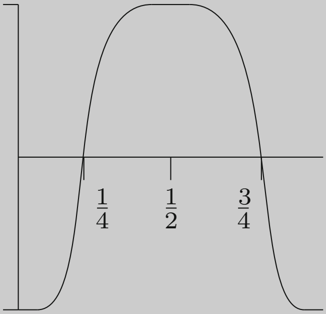
Multiple Nodes

From [2], we know that there is a set of multiple-node solutions.

(continued)



These are solutions which cross the x-axis multiple times. Currently, we hypothesize that the case for the symmetric multiple-node solutions will be similar to the symmetric one-node case. For instance, consider the case when the solution crosses at $x = \frac{1}{4}$ and $x = \frac{3}{4}$.



The solution on the interval $(0, \frac{1}{4})$ is symmetric (via $u(1-x) = -u(x)$) to the interval $(\frac{1}{4}, \frac{1}{2})$. Furthermore, the solution on the intervals $(\frac{1}{2}, \frac{3}{4})$ and $(\frac{3}{4}, 1)$ are also symmetric to the solution on the interval $(0, \frac{1}{4})$. Thus, it should follow from the proof of Theorem 1 with the exception that we define

$$u_0(x) = \begin{cases} -u_0(\frac{1}{2} - x) & \frac{1}{4} < x \leq \frac{1}{2} \\ -u_0(x - \frac{1}{2}) & \frac{1}{2} < x \leq \frac{3}{4} \\ u_0(1 - x) & \frac{3}{4} < x \leq 1. \end{cases}$$

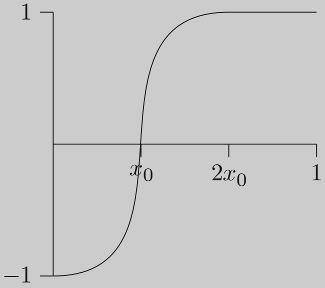
Notice that by similarly redefining $u_0(x)$ on specific intervals, we should be able to make a similar argument for other symmetric multiple node solutions as well.

(continued)

It would be nice if this was true. However, one aspect of the multiple-node case which must be accounted for is the local geometry of the functional J_ϵ for the multiple node solutions. From [1], we know that as the number of nodes increases, the dimension of the trough does as well. So, for the two-node solutions, we would be considering a three-dimensional trough which resembles a triangular plateau. This fact would be important to any proof of the multiple-node case.

Asymmetric Initial Values

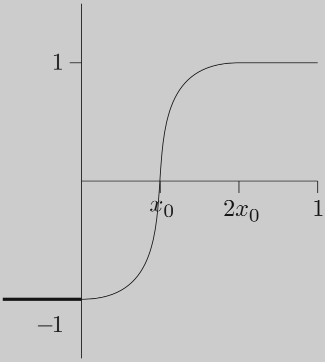
Finally, another question we wish to answer is whether there are asymmetric initial values $u_0(x)$ that are drawn to the base of the trough.



This is the case when one-node solutions cross the x-axis at some point $x_0 > \frac{1}{2}$ or $x_0 < \frac{1}{2}$. The initial attempt one might make with regard to this case is to make it into a symmetric one-node case. This can be done in two ways.

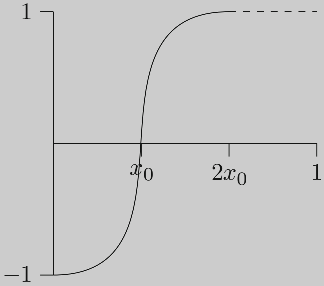
One idea is to extend the solution with a constant function over the interval $(-1 + 2x_0, 0)$.

(continued)



Note that we achieve the symmetry we desire; however, issues occur in that our problem only tells us what happens to the solution on the interval $(0, 1)$. There is no information to tell us what occurs to the solution on the interval $(-1 + 2x_0, 0)$. As such, this approach turns out to be nontrivial.

We can consider the other possibility by examining the solution on the interval $(0, 2x_0)$.



Again, we achieve the symmetry we desire; however, we must treat the section of the solution on the interval $(2x_0, 1)$ as a constant function which we will call $\eta(x, t)$. Our first issue with this approach arises when we try to use the existence and uniqueness theorem since the solution is only over the interval $(0, 2x_0)$. We must also consider what happens to $\eta(x, t)$ on the interval $(2x_0, 1)$ as time passes. While the symmetric part of the solution on $(0, 2x_0)$ converges to a solution of the problem by the symmetric argument we made before, we cannot guarantee that $\eta(x, t)$ will converge to the same solution. Hence, we must find a different approach to the asymmetric case.

While making the solution symmetric is an appealing way to check the stability of the asymmetric one-node solutions, as we saw, it is not a viable

(continued)

way to prove it. Therefore, we must consider a different approach. This new method involves considering a "curve" of initial conditions, call it $\gamma : [0,1] \longrightarrow C[0,1]$, where $\gamma(0)$ is an initial condition that goes to $u \equiv +1$ over time and $\gamma(1)$ is an initial condition that goes to $u \equiv -1$ over time. We construct γ so that $\gamma(t)$ is asymmetric for all t. We also construct $\gamma(t)$ so that any $\gamma(t)$ has energy below the two-node case as well as $u \equiv 0$. So $\gamma(t)$ can only be drawn to $u \equiv \pm 1$ or a one-node solution. Then we must examine these two cases. If $\gamma(t)$ is drawn to either $u \equiv +1$ or $u \equiv -1$ for all t, then we derive a contradiction. Since every initial condition goes somewhere, then $\gamma(t)$ must be drawn to a single-node solution for some t.

References

1. Drábek P , Manásevich R, Takáč P (2011) Manifolds of critical points in a quasilinear model for phase transitions. Contemp Math 540:95–134
2. Drábek P, Robinson S (2011) Continua of local minimizers in a non-smooth model of phase transitions. Zeitschrift für angewandte Mathematik und Physik ZAMP 62:609–622
3. Pao CV (1992) Nonlinear parabolic and elliptic equations. Plenum Press, New York

Two Fluid Flow in a Capillary Tube

Melissa Strait, Michael Shearer, Rachel Levy, Luis Cueto-Felgueroso, and Ruben Juanes

1 Model

The displacement of a fluid such as water by an injected finger of air in a narrow tube is a classic problem of fluid mechanics. Since the early experimental and theoretical work of Bretherton [4] and Taylor [10], there has been much research on the injection of one fluid into a different fluid resident in a thin tube [1, 3, 6, 7]. Characterizing such flows is significant not only for small scale fluid devices but also for modeling macroscopic two fluid flow in porous media [3, 9]. In this paper, we consider a recent model [5] that incorporates ideas from phase field theory, resulting in a fourth order nonlinear partial differential equation (PDE) similar to the PDE of thin liquid films [2]. The PDE possesses a spinodal-type instability at long wavelengths that we associate with the physical varicose or Plateau instability, in which the cylindrical gas finger, of sufficient length and for a range of widths, tends to break up into bubbles [7, 8].

We consider an axisymmetric flow of air displacing water in a cylindrical capillary tube. The dependent variable, which we refer to as the saturation u, is the cross-sectional area fraction of gas. The PDE model considered in [5] neglects the effect of gravity (which is reasonable for a thin tube, but can have a significant effect in wider tubes [7]) and takes the form

M. Strait • M. Shearer (✉)
Department of Mathematics, North Carolina State University, Raleigh, NC 27695, USA
e-mail: melissa.strait@gmail.com; shearer@ncsu.edu

R. Levy
Department of Mathematics, Harvey Mudd College, Claremont, CA 91711, USA

L. Cueto-Felgueroso • R. Juanes
Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

J. Rychtář et al. (eds.), *Collaborative Mathematics and Statistics Research*, Springer Proceedings in Mathematics & Statistics 109, DOI 10.1007/978-3-319-11125-4_14

$$\partial_t u + \partial_x f(u) = \partial_x \left(f(u)\lambda(u)\frac{1}{Ca}\partial_x \psi \right). \tag{1}$$

In this equation, the gas saturation $u = u(x,t)$ depends on x, the distance along the length of the tube and time t. The flux function f is the fractional flow rate, given by

$$f(u) = \frac{u}{u + k_w(u)},$$

and depends on the relative permeability $k_w(u)$ of water, which we take as either $k_w(u) = (1-u)^3$, or $k_w(u) = (1-u)^4$. We write $\lambda(u) = k_w(u)\frac{1}{M}(1 + (M-1)u)$, in which the mobility number $M = \eta_w/\eta_g > 1$ is the ratio of the viscosities η_w, η_g of the two fluids. The capillary number $Ca = U\eta_w/\gamma$ is the ratio of viscous and capillary forces, depending on U, a typical finger tip velocity, and γ, the surface tension between the two fluids. The function

$$\psi = C_1 g(u) - C_2\sqrt{\kappa(u)}\partial_x \left(\sqrt{\kappa(u)}\partial_x u \right),$$

is the chemical potential, derived as the variational derivative of a total free energy $F(u, \partial_x u)$. This has the form $F(u, \partial_x u) = C_1 F_0(u) + C_2\kappa(u)(\partial_x u)^2$, representing a bulk free energy plus an interface free energy. For simplicity in this paper, we take the bulk free energy to be a double-well quartic function of u, with

$$g(u) = u(1-u)(1-2u) = F_0'(u);$$

we generally take the coefficient of interfacial energy $\kappa(u)$, which is quadratic as $u \to 0$, to be $\kappa(u) = u^2$, as in [5]. The parameters C_1, C_2 are positive, and can be chosen so that the model accommodates the Young–Laplace law for the contact line at the tube entrance, where the gas finger attaches to the tube wall.

The objective of this paper is to outline a preliminary analysis of gas finger solutions of the PDE (1). These are traveling waves with the unusual property of being of finite extent, terminating at the tip of the gas finger in Fig. 1. Such traveling waves are solutions of a third order ordinary differential equation that is singular at the tip, where $u = 0$. With a change of variables, we transform the singular equation into a system that has a regular equilibrium at $u = 0$, and allows the numerical simulation of traveling waves. However, the solutions are not structurally stable, and depend on varying a parameter, specifically the finger width. Consequently, for each capillary number Ca in a specified range, there is a unique upstream width

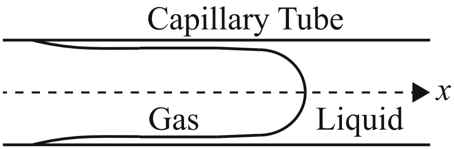


Fig. 1 Schematic of a gas finger displacing a liquid in a capillary tube

corresponding to a traveling wave. The range for Ca is determined by the nature of the equilibrium at $u = 0$, to avoid unphysical oscillations, since the saturation u has to remain non-negative. These properties are established in Sect. 2. In Sect. 3, we describe PDE simulations using a finite difference code, and compare the results to the traveling wave calculations and to experimental results of Taylor [10]. In the short Sect. 4, we describe the varicose instability by linearizing the PDE about a constant width finger. Finally, the results are discussed in Sect. 5.

2 Traveling Waves

In experiments, it is observed that the spherical tip of the gas finger travels with constant speed, and as the finger elongates, it leaves behind a nearly uniform layer of fluid adjacent to the tube wall [10]. To capture this behavior analytically, we seek traveling wave solutions $u(x,t) = u(x - st)$ of the PDE (1), where s is the wave speed. Such a solution has $u = 0$ at the tip of the gas finger. By translation invariance of the problem, we take this location to be $x = st$, without loss of generality. If $u_L > 0$ is the thickness of the fluid layer behind the tip, mathematically, the saturation should approach u_L as $\xi = x{-}st \to -\infty$. In summary, we have boundary conditions

$$u(-\infty) = u_L, \quad u(0) = 0. \tag{2}$$

Consistent with a smooth tip of the gas finger, we shall also assume that derivatives of u are bounded as $\xi \to 0$. Moreover, derivatives of $u(\xi)$ are taken to approach zero as $\xi \to -\infty$.

Substituting $u = u(\xi),\ \xi = x - st$ into (1) and integrating once, we obtain the third order ODE

$$K - su + f(u) = f(u)\lambda(u)\frac{1}{Ca}\psi', \quad \psi = C_1 g(u) - C_2\sqrt{\kappa(u)}\left(\sqrt{\kappa(u)}u'\right)',$$

where K is the constant of integration. Enforcing the boundary conditions at $\xi = 0, \xi = -\infty$, we find that $K = 0$, and that the speed s is given by the Rankine–Hugoniot condition

$$s = \frac{f(u_L)}{u_L}.$$

Incidentally, these conclusions depend on the degeneracy at $u = 0$, specifically that $f(0) = 0$. Now we have the ODE

$$\begin{aligned} -su + f(u) &= \frac{1}{Ca}C_1 H(u)\frac{dg(u)}{du}u' - C_2\frac{1}{Ca}H(u)\left(\sqrt{\kappa}(\sqrt{\kappa}u')'\right)', \\ H(u) &= f(u)\lambda(u). \end{aligned} \tag{3}$$

Equation (3) can be written as a first order system:

$$\begin{aligned}\sqrt{\kappa}u'(\xi) &= v\\ \sqrt{\kappa}v'(\xi) &= w\\ \sqrt{\kappa}w'(\xi) &= \frac{C_1}{C_2}G(u)v + \frac{Ca\sqrt{\kappa(u)}}{C_2H(u)}(su - f(u)),\end{aligned} \tag{4}$$

where $G(u) = \frac{dg(u)}{du}$. Since $\kappa(u) \sim u^2$ as $u \to 0$, system (4) has a singularity at $u = 0$. To remove the singularity, we introduce a new independent variable η. If $(u(\xi), v(\xi), w(\xi))$ is a traveling wave solution of (4), we set

$$\sqrt{\kappa(u(\xi))}\frac{d}{d\xi} = \frac{d}{d\eta},$$

and let $U(\eta) = u(\xi), V(\eta) = v(\xi), W(\eta) = w(\xi)$. For convenience, we revert to the lowercase letters, with $u(\eta)$, etc. Then, with $' = \frac{d}{d\eta}$,

$$\begin{aligned}u'(\eta) &= v\\ v'(\eta) &= w\\ w'(\eta) &= \frac{C_1}{C_2}G(u)v + \frac{Ca\sqrt{\kappa(u)}}{C_2H(u)}(su - f(u)).\end{aligned} \tag{5}$$

Now $H(u) = f(u)\lambda(u) \sim \frac{1}{M}u$, and $\sqrt{\kappa(u)} \sim u$, so the vector field represented by the right-hand side of Eq. (5) has a regular equilibrium at $u = 0$. Consequently, we seek trajectories $(u(\eta), v(\eta), w(\eta))$ from $(u_L, 0, 0)$ (as $\eta \to -\infty$), to $(0, 0, 0)$ (as $\eta \to +\infty$) with the property that u remains non-negative.

2.1 Equilibria

The system (5) has equilibria when $(u', v', w') = (0, 0, 0)$. Then $v = w = 0$, and equilibrium values of u are solutions of $su - f(u) = 0$, the intersection points of the flux function graph $y = f(u)$, and the line $y = su = \frac{f(u_L)}{u_L}u$. These curves necessarily intersect at $u = 0$ and $u = u_L$. Let u^* be defined as the value of u for which the tangent to the graph of f passes through the origin, shown in Fig. 2:

$$\frac{f(u^*)}{u^*} = f'(u^*).$$

A simple calculation shows that $u^* = 1 - 1/\sqrt{3}$. Let $s^* = (u^* + (1 - u^*)^3)^{-1}$ be the corresponding speed. For $u^* < u_L < 1$, there is a middle equilibrium u_M such that $0 < u_M \leq u^* < u_L$. Since $f(u) \sim u$ near $u = 0$, we observe that $u_M \to 0$ as $u_L \to 1$.

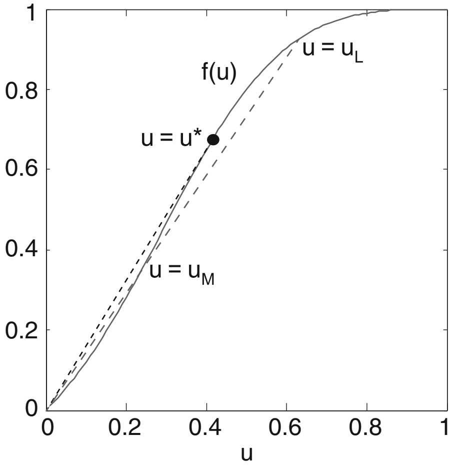


Fig. 2 Graph of the flux function $f(u)$, showing u^* and possible equilibrium values $u = u_L, u_M$ with the same $s = f(u_L)/u_L$

2.2 A Necessary Condition for Non-negative Traveling Waves

To obtain physically relevant solutions, in which the saturation u remains positive, we determine a bound on the quantity $M \cdot Ca$ by analyzing the linearized system at $(0, 0, 0)$. Recall that near $u = 0$,

$$H(u) = f(u)\lambda(u) = \frac{u(1-u)^3}{u+(1-u)^3}\frac{1}{M}(1+(M-1)u) \sim \frac{1}{M}u,$$

$$f'(0) = 1, \quad \text{and} \quad G(0) = 1.$$

Therefore, system (5) linearized around $u = v = w = 0$ has the structure

$$\begin{bmatrix} u' \\ v' \\ w' \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{MCa}{C_2}(s-1) & \frac{C_1}{C_2} & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix}.$$

The nature of the equilibrium at the origin is determined by the eigenvalues $\lambda_k,\ k = 1, 2, 3$ of the coefficient matrix. These are the three roots of the function

$$y(\lambda) = \lambda^3 - \frac{C_1}{C_2}\lambda - \frac{Ca}{C_2}M(s-1). \tag{6}$$

Note that the $\lambda_1\lambda_2\lambda_3 = \frac{Ca}{C_2}M(s-1) > 0$, and $\lambda_1 + \lambda_2 + \lambda_3 = 0$. Consequently, one eigenvalue is positive and the other two are either negative or are complex conjugates and have negative real parts. The latter eigenvalues correspond to the two-dimensional stable manifold of the equilibrium at the origin, on which the desired trajectory must lie. In order to prevent the gas saturation, u on this manifold from becoming negative, all three eigenvalues must be real, since otherwise solutions will have oscillations around $u = 0$, and u will not remain positive.

To determine the range of parameters for which all three eigenvalues are real, we analyze the function $y(\lambda)$ in (6). The local maximum, $y(\lambda_m)$, occurs at $\lambda_m = -\sqrt{\frac{C_1}{3C_2}}$.

There are three real roots when $y(\lambda_m) > 0$, leading to the following lemma.

Lemma 1 *Suppose there is a traveling wave solution of* (1)*, satisfying* (2) *with* $u \geq 0$.

(a) *Then*

$$M \cdot Ca < \frac{2}{3\sqrt{3}(s-1)}\sqrt{\frac{C_1^3}{C_2}}, \tag{7}$$

where $s = f(u_L)/u_L$.

(b) *Suppose moreover, that* $\overline{s} > 1$ *is defined by*

$$M \cdot Ca = \frac{2}{3\sqrt{3}(\overline{s}-1)}\sqrt{\frac{C_1^3}{C_2}}.$$

Then $s < \min(\overline{s}, s^*)$.

The implication of part (b) is that if $\overline{s} < s^*$, then the possible range of values of the traveling wave speed s is restricted, and consequently, the possible values of u_L are also restricted. Specifically, let $\overline{u}_L$ be defined by $\overline{s} = f(\overline{u}_L)/\overline{u}_L$. Then in order that $1 < s < \overline{s}$, we must have $\overline{u}_L < u_L < 1$.

2.3 The Equilibrium at $u_L > 0$

Since $H(u_L) > 0$, the equilibrium at $u = u_L$ is regular, and the Jacobian of F is given by

$$DF(u_L, 0, 0) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{Ca\sqrt{\kappa(u_L)}}{C_2 H(u_L)}(s - f'(u_L)) & \frac{C_1}{C_2}G(u_L) & 0 \end{bmatrix}.$$

The characteristic polynomial associated with this system is

$$y(\lambda) = \lambda^3 - A\lambda - B,$$

where $A = \frac{C_1}{C_2}G(u_L)$, and $B = \frac{u_L Ca}{C_2 H(u_L)}(s - f'(u_L))$. The eigenvalues, given by the zeroes of $y(\lambda)$, vary continuously with the coefficients A, B. For $A = 0$, the three eigenvalues are (complex) cube roots of B. Consequently, if $u_L > u^*$, then $B > 0$, and there is one positive real eigenvalue, and a pair of complex conjugate eigenvalues with negative real parts.

If an eigenvalue crosses the imaginary axis as A is varied, then for some A, the real part of the eigenvalue vanishes, so $\lambda = i\beta$, $\beta \in \mathbb{R}$. Therefore,

$$y(\lambda) = -i\beta^3 - Ai\beta - B = 0,$$

a contradiction. We conclude that, for $u_L > u^*$, two eigenvalues of the equilibrium at u_L have negative real parts, and the third eigenvalue is real and positive. Consequently, the local dynamics are described by a two-dimensional stable manifold and a one-dimensional unstable manifold at u_L. Similarly, if $0 < u_L < u^*$, the equilibrium at u_L has a two-dimensional unstable manifold and a one-dimensional stable manifold, since in that case, we have $s < f'(u_L)$ and $B < 0$.

Finally, we observe from the structure of $DF(u)$ that right eigenvectors have the form $(1, \lambda, \lambda^2)^T$, for each eigenvalue λ of $DF(u)$.

2.4 *Computing the Traveling Wave Solutions*

We seek a solution of system (5) that connects $(u_L, 0, 0)$ to $(0, 0, 0)$ with $u_L > u^*$. Such a solution corresponds to a trajectory that leaves $(u_L, 0, 0)$ on its one-dimensional unstable manifold $W^U(u_L)$, and intersects the two-dimensional stable manifold $W^S(0)$ of the equilibrium at $u = 0$. Then the entire trajectory lies in $W^S(0)$. However, this intersection has to be achieved by varying a parameter, suggesting a shooting method. Geometrically, the intersection is codimension one. In this sense, the corresponding traveling wave solutions of (1) are undercompressive, as discussed in [2].

Let the parameters Ca and M be fixed. We use an ODE solver in MATLAB to approximate the trajectory leaving $(u_L, 0, 0)$ along $W^U(u_L)$, with $u(\eta)$ decreasing. To this end, we initiate the ODE solver by taking (u, v, w) a small distance $\epsilon > 0$ away from $(u_L, 0, 0)$ along the eigenvector $-(1, \lambda, \lambda^2)$, where λ is the positive eigenvalue associated with the equilibrium at u_L :

$$(u, v, w)(0) = (u_L, 0, 0) - \epsilon(1, \lambda, \lambda^2).$$

We solve the system (5) in MATLAB, and track the sign of $u(\eta)$ and $u'(\eta)$ for each choice of u_L. In extreme cases, the trajectory exhibits contrasting behavior, corresponding to missing $W^S(0)$ on one side or the other: (a) For u_L close to $u = 1$, $u(\eta)$ becomes negative, and (b) for u_L close to u^*, $u(\eta)$ remains positive but has a positive minimum before exceeding $u = u_L$. These two behaviors are incorporated

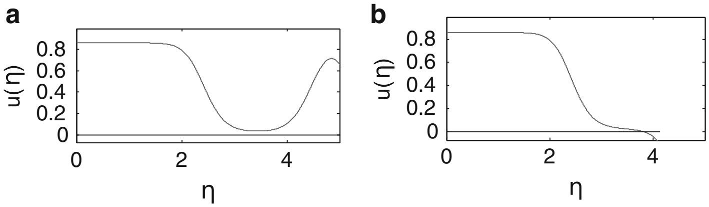


Fig. 3 Trajectories exhibiting the two behaviors seen when using the bisection method to solve system (5). (**a**) $u(\eta) > 0$ has a minimum. (**b**) $u(\eta)$ crosses $u = 0$

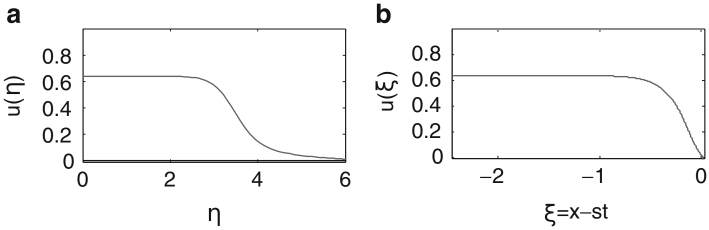


Fig. 4 Traveling waves: (**a**) in the transformed variable η; (**b**) in the physical variable ξ

into an interval division algorithm (bisection method) to approximate the value of u_L for which $u(\eta)$ remains positive, while its minimum is pushed off towards $\eta = \infty$.

Examples of trajectories with the two behaviors are shown in Fig. 3, and a typical trajectory $u = u(\eta)$ is shown in Fig. 4a.

As we vary the capillary number Ca, we find new values of $u_L = u_L(Ca)$ for which there is a trajectory from u_L to $u = 0$. A plot of $1 - u_L$ against Ca is shown in Fig. 6, together with comparisons to experiment and PDE simulations, as explained below.

2.5 Inverting the Transformation

The trajectories in the previous subsection were obtained in the transformed independent variable η, with a solution $\tilde{u}(\eta)$, $-\infty < \eta < \infty$. To convert back to the physical variable ξ, which remains finite, and derive the desired function $u(\xi)$, we first recall that the change of variables from ξ to η was predicated on the existence of a solution $u(\xi)$, so that

$$\sqrt{\kappa(u)}\frac{du}{d\xi} = \frac{d\tilde{u}}{d\eta}. \tag{8}$$

However, if the change of variables is $\eta = \eta(\xi)$, we have $\tilde{u}(\eta(\xi)) = u(\xi)$. Inverting, if $\xi = \xi(\eta)$ then from the chain rule

$$\frac{d\xi}{d\eta} = u(\xi(\eta)).$$

Solving this ODE using separation of variables gives

$$\xi = -\int_{\eta}^{\infty} \tilde{u}(\bar{\eta}) d\bar{\eta}.$$

Since we assume our traveling wave solution $u(\eta) \approx 0$ for $\eta \geq N$, for large enough $N > 0$, then

$$\xi = -\int_{\eta}^{N} \tilde{u}(\eta) d\eta.$$

The traveling wave solution with the inversion completed is shown in Fig. 4b, where we have used $\kappa(u) = u^2$ for both the ODE solver and the transformation.

2.6 Conclusion

For each $Ca \in [10^{-3}, 1]$ and for fixed C_1, C_2, and M satisfying (7), the method described in Sect. 2.4 generates a unique traveling wave solution to (1) subject to (2). We conjecture, and find numerically, that for each Ca and fixed C_1, C_2, and M there is a unique u_L with a traveling wave connecting u_L to 0. The numerical method for finding u_L for each value of Ca is robust, but quite sensitive, meaning that u_L has to be calculated to a large number of decimal places (around 12–14) in order to have the flat portion near $u = 0$ extend as in Fig. 4a for example.

3 PDE Simulations

The PDE (1) is solved using an implicit finite difference method to model the injection of a gas finger into a fluid filled tube. A fixed domain, $x \in [-L, L]$, is used with boundary conditions

$$u(-L,t) = 1, \quad u(L,t) = 0, \quad u'(-L,t) = 0, \quad u'(L,t) = 0,$$

L is chosen to be large enough to assume zero gas saturation at $x = L$. Finite difference simulations in Fig. 5 show a traveling wave advancing ahead of a rarefaction wave, connected by a plateau region of residual fluid.

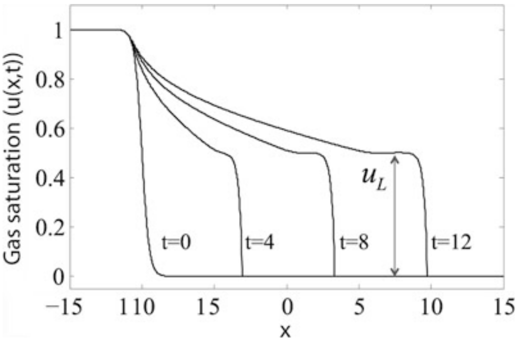


Fig. 5 Finite difference simulations of air injection with $L = 15$, $Ca = 0.5$, $M = 10$, $C_1 = 0.2$, and $C_2 = \frac{1}{7}$

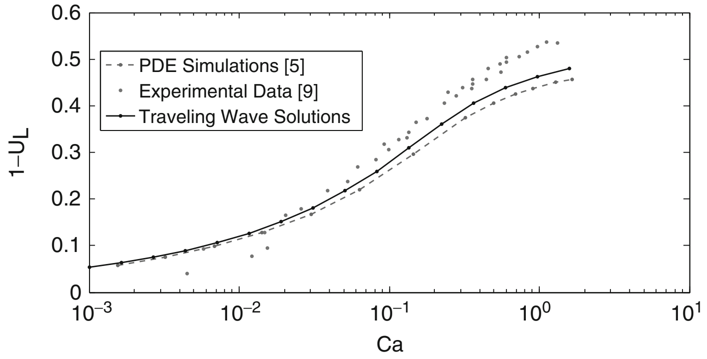


Fig. 6 Comparison of the residual fluid remaining, $1 - u_L$, in simulations and experiments [10]

The height of the traveling wave from PDE simulations can be compared with the traveling wave height computed from the ODE (4) and results from classical experiments by Taylor. The PDE and ODE simulation data along with experimental data from [10] are shown in Fig. 6. The amount of fluid left behind, between the gas finger and tube wall, $1 - u_L$, is plotted against the capillary number, Ca. Both simulations and experiments show that as Ca increases, the amount of fluid left increases. The PDE and ODE simulations closely agree for $Ca \in [10^{-3}, 1.5]$. Both simulations predict the same trend as the experimental data [10], but under-predict the amount of fluid remaining for large capillary numbers.

The model (1) assumes the relative permeability of the water, λ_w, has the form $\lambda_w = (1 - u)^3$. The agreement between model simulations and experimental data for larger Ca, in the range $[10^{-1}, 1.5]$, can be improved by changing the form of λ_w to a quartic function, $\lambda_w = (1 - u)^4$. This changes $f(u)$ and $\lambda(u)$ in (1) to

$$f(u) = \frac{u}{u + (1-u)^4}, \qquad \lambda(u) = (1-u)^4 \frac{1}{M}(1 + (M-1)u).$$

Refining the choice of C_1 and C_2 can also better fit the simulations to experimental results.

4 Varicose Instability

The Rayleigh–Plateau instability causes long gas fingers to break into bubbles, as observed in experiments in [7]. To compare this physical instability to the stability of the PDE we analyze (1) linearized about a constant u_0. For $u(x,t) = u_0 + \epsilon\tilde{u}(x,t)$, with $\epsilon << 1$, the linearized PDE is

$$\tilde{u}_t + f'(u_0)\tilde{u}_x = \frac{H(u_0)}{Ca}\left(C_1 g'(u_0)\tilde{u}_{xx} - C_2 u_0^2 \tilde{u}_{xxxx}\right). \tag{9}$$

To find the dispersion relation between the frequency, $\omega(\xi)$, and wave number ξ, we assume a perturbation of the form $\tilde{u}(x,t) = e^{i(\xi x + \omega t)}$ and substitute into (9) resulting in

$$\omega(\xi) = -f'(u_0)\xi + i\,\frac{H(u_0)}{Ca}\left[C_1 g'(u_0)\xi^2 + C_2 u_0^2 \xi^4\right]. \tag{10}$$

The perturbation decays with time if and only if Im $\omega > 0$. This results in the stability restriction

$$0 < \frac{H(u_0)}{Ca}\left[C_1 g'(u_0)\xi^2 + C_2 u_0^2 \xi^4\right]. \tag{11}$$

In order for perturbations to decay for all wave numbers, ξ, $g'(u_0)$ must be positive. However, for the choice of nonlinearity in this paper, $g'(u) < 0$ in the range $0.212 < u < 0.788$. Thus, the solutions can be expected to develop long wave instabilities in this range.

We can also determine the wavelengths that increase in amplitude in this range of u. The wavelength, λ, is related to the wave number, ξ, by the

$$\lambda = \frac{2\pi}{\xi}.$$

To find the range of unstable λ, we determine the values of ξ such that (11) is not satisfied. At $u = 0.5$, $g'(u)$ attains its minimum value, $g(0.5) = -0.5$. For definiteness, let $C_1 = 1$, $C_2 = 0.1$, values used in the simulations. Then $\operatorname{Im}\omega < 0$ when $|\xi| \leq \sqrt{20}$. Therefore the range of unstable wavelengths in this particular case is $\lambda \geq \frac{2\pi}{\sqrt{20}} \approx 1.4$.

5 Discussion

In this paper, we have verified that the phase field model of [5] captures the structure of a gas finger being forced into a capillary tube filled with water. The model

PDE describes the evolution of the gas saturation assuming an axisymmetric cross-section through the gas and water. It incorporates a bulk free energy and an interfacial energy, with surface tension incorporated into a capillary number Ca. The structure of the PDE solution is approximately a spreading rarefaction wave attached to the tube entrance, preceded by the finger, which is a traveling wave that terminates at the finger tip. To calculate the traveling wave, we use a change of variables which sends the tip to infinity and makes the zero saturation limit at the tip a regular equilibrium for the associated ODE system.

The ODE solution has to remain positive in order to be physical, and this entails a maximum capillary number. Below this threshold, we calculate a value of the finger width (more precisely the saturation u_L) in the traveling wave using a shooting method. The structure of the wave is similar to that observed in driven thin liquid films, also modeled with a fourth order PDE [2]. The values of u_L compare well with those observed in finite difference simulations of the PDE, and with experimental observations of Taylor [10].

In all of these comparisons, simple constitutive functions have been used, specifically, $g(u) = u(u-1)(1-2u)$, and $\kappa(u) = u^2$. The varicose instability, generated as a result of the non-monotonicity of $g(u)$, can be tuned using different functions $g(u)$, and calibrated against the range of finger widths at which the instability is observed experimentally. The function $\kappa(u)$ should admit a spherical cap at the finger tip. It is reasonable to choose this function so that stationary bubbles attached to the tube wall are solutions of the PDE. At zero contact angle, this requires $\kappa(u) = cu^2(1-u)$, with $c > 0$ depending on the form of $g(u)$. At other contact angles, there is a corresponding formula. These changes in the constitutive laws are easily incorporated into both the ODE and PDE solvers, and will be reported on in the future.

Acknowledgements Research of the first two authors was funded by NSF grant DMS 0968258. Research of the third author was funded by NSF grant 0968154. Research of the last two authors was funded by a DOE CAREER Award, DE-SC0003907, and a DOE Mathematical Multifaceted Integrated Capability Center, DE-SC0009286.

References

1. Beresnev I, Gaul W, Vigil RD (2011) Thickness of residual wetting film in liquid–liquid displacement. Phys Rev E 84:026327
2. Bertozzi AE, Münch A, Shearer M (1999) Undercompressive shocks in thin film flows. Physica D 134(4):431–464
3. Blake TD, De Coninck J (2004) The influence of pore wettability on the dynamics of imbibition and drainage. Colloids Surf A 250(1–3):395–402
4. Bretherton FP (1961) The motion of long bubbles in tubes. J Fluid Mech 10:166–188
5. Cueto-Felgueroso L, Juanes R (2012) Macroscopic phase-field model of partial wetting: bubbles in a capillary tube. Phys Rev Lett 108:144502
6. De Lózar A, Juel A, Hazel AL (2008) The steady propagation of an air finger into a rectangular tube. J Fluid Mech 614:173–195

7. Duclaux V, Clanet C, Quéré D (2006). The effects of gravity on the capillary instability in tubes. J Fluid Mech 556:217–226
8. Goren SL (1962) The instability of an annular thread of fluid. J Fluid Mech 12:309–319
9. Scheidegger AE (1974) The physics of flow through porous media. University of Toronto Press, Toronto
10. Taylor GI (1961) Deposition of a viscous fluid on the wall of a tube. J Fluid Mech 10:161–165