

# involve

a journal of mathematics

## Editorial Board

Kenneth S. Berenhaut, *Managing Editor*

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



# involve

msp.org/involve

## EDITORS

### MANAGING EDITOR

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

### BOARD OF EDITORS

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

## PRODUCTION


Silvio Levy, Scientific Editor

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2014 is US \$120/year for the electronic version, and \$165/year (+\$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

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

Involve peer review and production are managed by EditFLOW<sup>®</sup> from Mathematical Sciences Publishers.

PUBLISHED BY

 **mathematical sciences publishers**  
nonprofit scientific publishing

<http://msp.org/>

© 2014 Mathematical Sciences Publishers

# A median estimator for three-dimensional rotation data

Melissa Bingham and Zachary Fischer

(Communicated by Michael Dorff)

The median is a way of measuring the center of a set of data that is robust to outlying values. However, the concept of a median for three-dimensional rotation data has been largely nonexistent. Although there are already ways to measure the center of three-dimensional rotation data using the idea of a “mean rotation”, the median estimator developed here is shown to be less influenced by outlying data points. A simulation study that investigates scenarios under which the median is an improvement over the mean will be discussed. An application to a three-dimensional data set in the area of human motion will be considered.

## 1. Introduction

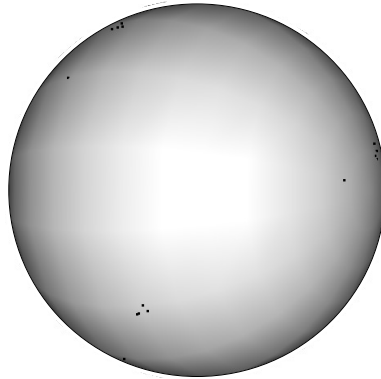
Data in the form of three-dimensional rotations are common in the areas of human motion and biomechanics, since they can be used to characterize the relative orientation of one body segment with respect to another during movement. We use data collected in a study by Rancourt, Rivest, and Asselin [Rancourt et al. 2000] to motivate the need for a median for this type of data. During the study, individuals drilled into six locations on a vertical panel, with each subject repeating the drilling five times. Infrared emitting diodes placed on the subject’s hand, forearm, arm, and torso allowed for collection of orientations of the wrist, elbow, and shoulder during the drillings. Figure 1 shows five repeated wrist orientations for the drillings performed by one of the subjects studied. Since each observation is a three-dimensional rotation, it can be represented mathematically as a  $3 \times 3$  orthogonal rotation matrix with determinant 1 (i.e., is a member of the rotation group  $SO(3)$ ) and can be displayed graphically as a set of three points on the sphere, corresponding to the locations of three orthogonal axes. Notice that one of the five orientations seems to be an outlying value, as it is not clustered near the other four observations.

---

*MSC2010:* 62H11, 62P99.

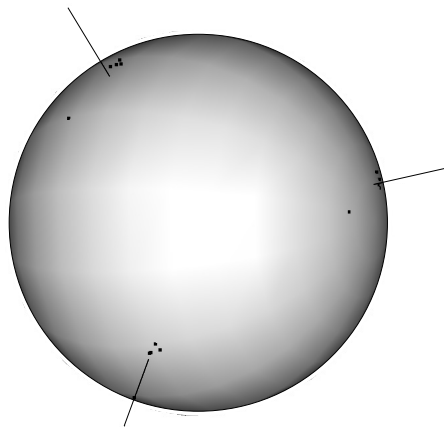
*Keywords:* directional statistics, rotations, median.

This research was supported by NSF grant DMS-1104409.

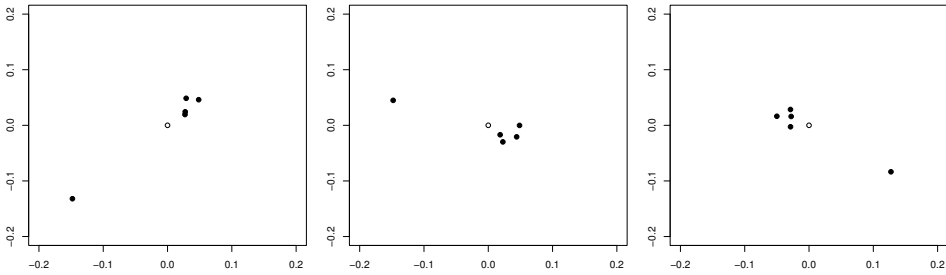


**Figure 1.** Five repeated wrist orientations from the drilling study (represented as a set of three points).

A common first step in data analysis is to characterize data according to some measure of center, and we attempt to do so here for the repeated drilling data. First consider using the “mean rotation” that is commonly used as a measure of center for three-dimensional rotation data [León et al. 2006; Bingham et al. 2009; Khatri and Mardia 1977]. If  $\mathbf{O}_1, \dots, \mathbf{O}_n \in \text{SO}(3)$  is a random sample of three-dimensional rotations and  $\bar{\mathbf{O}} = \sum_{i=1}^n \mathbf{O}_i$ , the mean rotation is defined as  $\mathbf{T} = \mathbf{V}\mathbf{W} \in \text{SO}(3)$ , where  $\bar{\mathbf{O}} = \mathbf{V}\mathbf{\Sigma}\mathbf{W}$  is the singular value decomposition of  $\bar{\mathbf{O}}$ . The singular value decomposition is necessary since  $\bar{\mathbf{O}}$  itself is not necessarily an element of  $\text{SO}(3)$  and therefore cannot serve as a mean rotation. The mean rotation of the five wrist orientations shown in Figure 1 was found and is displayed as the set of three axes in Figure 2.



**Figure 2.** Five repeated wrist orientations from the drilling study (represented as a set of three points) and the mean rotation (represented as a set of three perpendicular axes).



**Figure 3.** Stereographic projections of the five repeated wrist orientations displayed in Figure 2, with the center point representing the mean.

Figure 3 shows the same data as a stereographic projection, with the open circle at the center corresponding to the position of the mean rotation. It can be seen that the mean is not robust to outliers, as it is pulled towards the one observation that might be considered an outlier. In cases like this, a median would be preferred as a measure of center due to its robustness. While the median is typically thought of as the value that divides an ordered distribution in half, this definition is only easily applied to data that exhibit some type of natural ordering, and this property is nonexistent for three-dimensional rotation data. Therefore, we propose a possible median estimator for three-dimensional rotations in Section 2. In Section 3 we examine the effectiveness of this median through a simulation study, and in Section 4 we revisit the wrist orientations to show that the median provides a better measure of center for this data.

### 2. Development of a median estimator

Because three-dimensional rotation data do not have a natural ordering, we cannot simply define the median as the value that divides the ordered distribution in half. Instead we will consider the optimality property of the median when developing a possible median estimator for such data. The optimality property tells us that the mean absolute deviation attains a minimum when the deviation is measured from the median [Lee 1995], so that for a random variable  $X$ ,  $E|X - m|$  is minimized where  $m = \text{median}$ .

To use the optimality property for three-dimensional rotations, we need a concept of “distance” (or deviation) between two orientations. For  $\mathbf{O}_1, \mathbf{O}_2 \in \text{SO}(3)$ , there exists a vector  $\mathbf{U} \in \mathbb{R}^3$  and an angle  $r \in [0, \pi]$  such that a rotation of  $\mathbf{O}_1$  about  $\mathbf{U}$  by  $r$  results in  $\mathbf{O}_2$ . The angle  $r$  is sometimes referred to as a *misorientation angle* [Morawiec 2004] and we consider this angle as a measure of distance between rotations  $\mathbf{O}_1$  and  $\mathbf{O}_2$ . Suppose  $\mathbf{P}_1, \dots, \mathbf{P}_n$  is a set of  $n$  rotations in  $\text{SO}(3)$ , and

let  $r(\mathbf{P}_i, \mathbf{M})$  denote the misorientation angle between  $\mathbf{P}_i$  and  $\mathbf{M}$ . We define the median rotation as the element of  $\text{SO}(3)$  that minimizes  $f(\mathbf{M}) = \sum_{i=1}^n r(\mathbf{P}_i, \mathbf{M})$ , so that the average deviation from all data points would be minimized by the median rotation. In the next section we compare this median rotation to the mean rotation introduced in Section 1 through a simulation study.

### 3. Simulation study

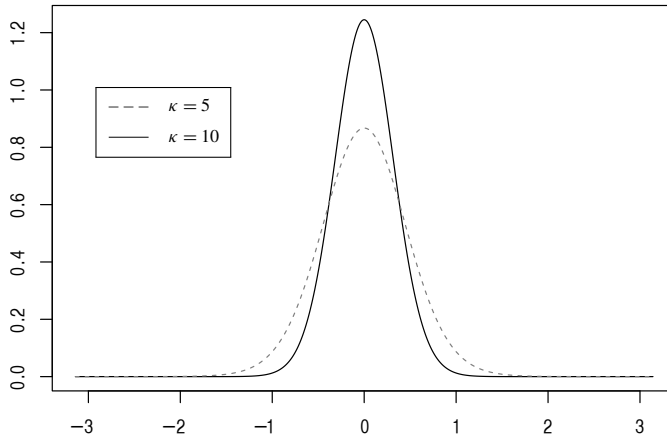
To examine the effectiveness of the median estimator defined in Section 2, we simulate random rotations from the uniform axis-random spin (UARS) distributions of [Bingham et al. 2009] under various conditions. The UARS distributions are a symmetric class of distributions for three-dimensional rotations, and Bingham et al. [2009] begin their development of this class by discussing a technique for data simulation. Simulation begins by starting with the  $3 \times 3$  identity matrix, denoted by  $\mathbf{I}_{3 \times 3}$ . Then a unit vector  $\mathbf{U}$  that is uniformly distributed on the  $\mathbb{R}^3$ -sphere is generated. Next, the angle  $\theta \in (-\pi, \pi]$  is independently generated from a circular distribution that is symmetric about 0 and depends on a concentration parameter  $\kappa$  (with density  $C(\theta | \kappa)$ ). Rotating  $\mathbf{I}_{3 \times 3}$  around  $\mathbf{U}$  by the angle  $\theta$  results in an orientation  $\mathbf{P}$ . If this process is repeated  $n$  times, we arrive at a set of orientations  $\mathbf{P}_1, \dots, \mathbf{P}_n$  that is scattered about the center  $\mathbf{I}_{3 \times 3}$ . Since  $\kappa$  controls the angle  $\theta$  that is generated from the circular distribution, it also controls the spread of the resulting orientations from their center. Now, by letting  $\mathbf{M}_i = \mathbf{S}\mathbf{P}_i$  for  $i = 1, \dots, n$ , the rotations  $\mathbf{M}_1, \dots, \mathbf{M}_n$  have center at  $\mathbf{S}$ . The rotation  $\mathbf{M}_i$  is said to have UARS distribution with parameters  $\mathbf{S}$  (indicating center) and  $\kappa$  (indicating spread), which is denoted by  $\mathbf{M}_i \sim \text{UARS}(\mathbf{S}, \kappa)$ .

For the simulation study considered here, we will use the UARS class with  $\theta$  coming from the von Mises circular distribution with mean 0. The von Mises distribution is the most commonly used circular distribution because it is symmetric and unimodal, and as  $\kappa \rightarrow \infty$ , the distribution approaches the normal distribution with standard deviation  $1/\kappa$ . The density for  $\theta$  is

$$C(\theta | \kappa) = [2\pi I_0(\kappa)]^{-1} \exp[\kappa \cos(\theta)], \quad \theta \in (-\pi, \pi],$$

where  $I_0(\kappa)$  is the modified Bessel function of order zero. (For more on the von Mises distribution see [Mardia and Jupp 2000].) See Figure 4 for a plot of the von Mises density for  $\kappa = 5$  and  $\kappa = 10$ , which shows how the concentration parameter  $\kappa$  affects the spread of the distribution. We refer to the von Mises version of the UARS class as vM-UARS.

For the simulation study considered here, we generated  $\mathbf{S}_1$  and  $\mathbf{S}_2$  uniformly in  $\text{SO}(3)$ . A total of  $n$  rotations,  $\mathbf{Q}_1, \dots, \mathbf{Q}_n$ , were simulated from the vM-UARS



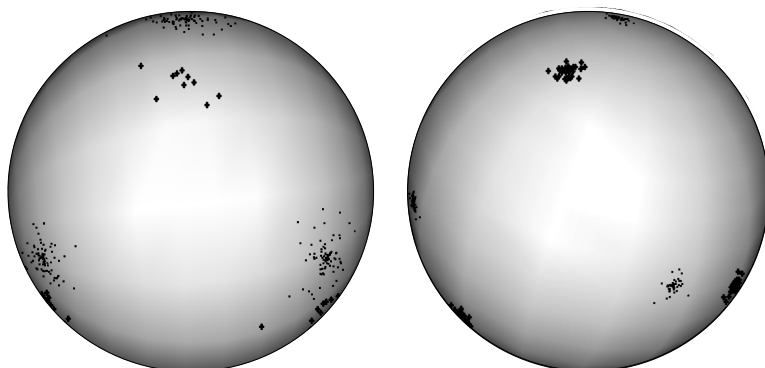
**Figure 4.** Von Mises circular density for concentration parameter  $\kappa = 5, 10$ .

distribution, with a proportion  $p$  being  $\text{vM-UARS}(\mathcal{S}_1, \kappa)$  and  $1 - p$  being  $\text{vM-UARS}(\mathcal{S}_2, \kappa)$ . We think of this data set as being composed of  $100p\%$  “outliers”, so that the data is centered at  $\mathcal{S}_2$  with outliers near  $\mathcal{S}_1$ . We then found the misorientation angle between  $\mathcal{S}_1$  and  $\mathcal{S}_2$ , called  $V$  (i.e.,  $V = r(\mathcal{S}_1, \mathcal{S}_2)$ ). We think of  $V$  as measuring the distance between the center of the data and where the outliers are located. The values of  $\kappa$  considered in the simulation study were 5, 10, 50, 100, and 500. The values of  $n$  considered were 10, 50, 100, and 500. The values of  $p$  used for each choice of  $n$  are given in Table 1. Figure 5 shows a plot of  $\mathcal{Q}_1, \dots, \mathcal{Q}_{100}$  on the sphere for two different cases of  $\kappa, p$ , and  $V$ . In both instances, the proportion of bolder, cross-shaped points is  $p$  (representing the “outliers”).

Once  $\mathcal{Q}_1, \dots, \mathcal{Q}_n$  were generated, the mean and median rotations, referred to as  $\mathbf{N}$  and  $\mathbf{M}$ , respectively, were found. To measure the “distance” from the mean to the simulated rotations, we considered the sum of the misorientation angles  $\sum_{i=1}^n r(\mathcal{Q}_i, \mathbf{N})$ . A similar measure,  $\sum_{i=1}^n r(\mathcal{Q}_i, \mathbf{M})$ , was found for the median. We compared the mean and median by considering the difference of these distances,

	Choices for $p$
$n = 10$	0.1, 0.3, 0.5
$n = 50$	0.04, 0.1, 0.3, 0.5
$n = 100$	0.01, 0.05, 0.1, 0.5
$n = 500$	0.002, 0.01, 0.1, 0.5

**Table 1.** Choices of  $p$  (proportion of “outliers”) used in the simulation study for each value of  $n$  considered.



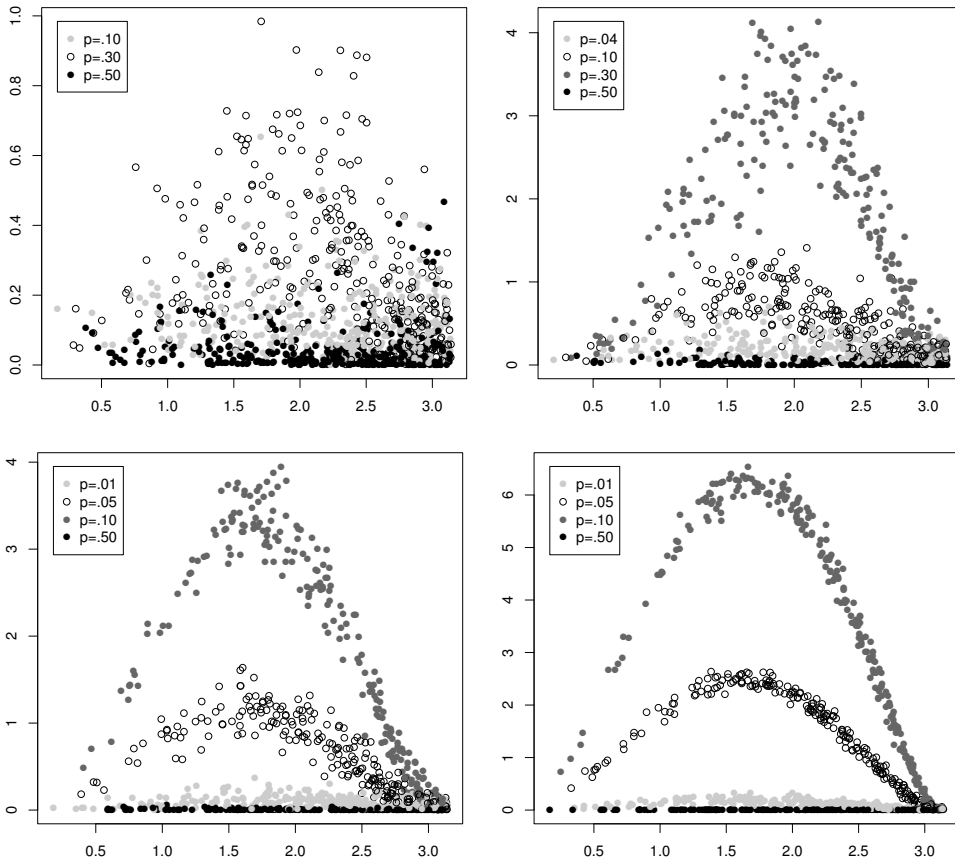
**Figure 5.** 100 simulated orientations with (left)  $\kappa = 50$ ,  $p = 0.10$ , and  $V = 0.488$  and (right)  $\kappa = 500$ ,  $p = 0.50$ , and  $V = 1.092$ .

$R(\mathbf{M}, \mathbf{N}) = \sum_{i=1}^n r(\mathbf{Q}_i, \mathbf{N}) - \sum_{i=1}^n r(\mathbf{Q}_i, \mathbf{M})$ . Note that larger values of  $R(\mathbf{M}, \mathbf{N})$  indicate that the median rotation is outperforming the mean rotation.

For each combination of  $\kappa$  and  $n$ , 1000 rotation data sets were generated. For each data set,  $p$  was chosen randomly from the possible values listed in Table 1. A plot was then created with  $V$ , the “distance” between the uniformly selected  $\mathbf{S}_1$  and  $\mathbf{S}_2$ , on the horizontal axis and  $R = R(\mathbf{M}, \mathbf{N})$  on the vertical axis, with different plot characters and shades of gray used to represent the various values for  $p$ . Although a total of 20 plots were made (one for each combination of  $\kappa$  and  $n$ ), only a few, which show the general relationships seen in all plots, are provided here.

Figure 6 contains the plots for  $(\kappa, n)$  combinations of  $(5, 10)$ ,  $(10, 50)$ ,  $(50, 100)$ , and  $(500, 100)$ . As expected, when  $\kappa$  increases (meaning the simulated data is less spread) or  $n$  increases, the relationship between  $V$  and  $R$  becomes more defined. An interesting and unexpected feature seen in the plots is the quadratic-type relationship between  $V$  and  $R$ . We see that the maximum values of  $R$ , which coincide with the median most outperforming the mean, happen in the middle of the range of  $V$  values. One might expect that  $R$  would increase as  $V$  increases and the outliers move farther from the rest of the data. Instead, when the outliers are at a misorientation angle of  $\pi$  away, the mean and median are almost identical. This phenomenon is due to the fact that the three axes are orthogonal, making it impossible for them to be simultaneously pulled toward the outliers. Figure 7 contains 100 orientations plotted as a set of three points on the sphere, of which 10 would be considered outliers ( $p = 0.10$ ). The points around the  $x$ -,  $y$ -, and  $z$ -axes have been plotted using three different colors so that it is clear which outliers belong to which cluster of points. The angle between the cluster of points of seven points and the three outliers is  $V = \pi$ . The mean and median are indistinguishable from one another on this plot, and both are represented by the set of three axes. If one axis were to be

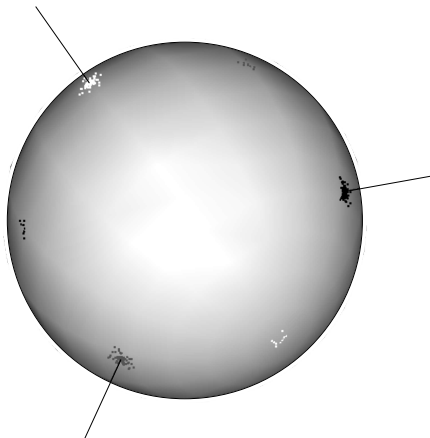




**Figure 6.** Plot of  $R$  (vertical axis) against  $V$  (horizontal axis) for 1000 simulated data sets using (top left)  $\kappa = 5, n = 10$ , (top right)  $\kappa = 10, n = 50$ , (bottom left)  $\kappa = 50, n = 100$ , and (bottom right)  $\kappa = 500, n = 100$ .

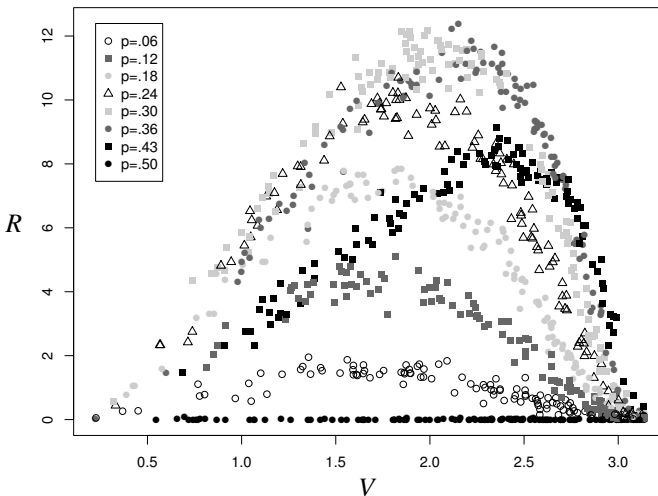
pulled toward the outliers, the other axes would not be able to be pulled toward the outliers and still remain orthogonal. As a result, the mean is not influenced by the outliers in this case. Therefore, the quadratic-type relationship between  $V$  and  $R$ , while unexpected, is understandable after considering the orthogonality of the axes within a three-dimensional rotation data point.

We can also discuss the plots in regards to the proportion of outliers,  $p$ . In all plots we see that with  $p = 0.50$  the mean and median are generally equivalent ( $R$  near 0). This is due to the fact that with an equal number of data points coming from the two centers  $S_1$  and  $S_2$ , both the mean and median will tend to be half-way between these centers (with neither measure experiencing more pull toward one center). From the plots presented in Figure 6 it appears that, with the exception

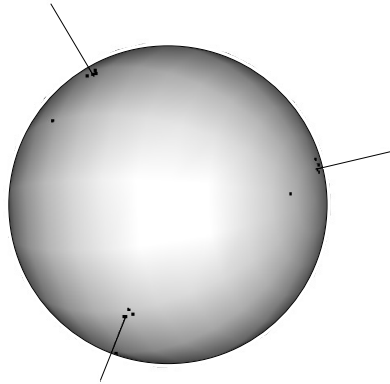


**Figure 7.** Plot of 100 orientations, represented as a set of three points on the sphere. Here  $p = 0.10$ ,  $V = \pi$ , and the set of three axes represents the mean/median.

of  $p = 0.50$ , the value of  $R$  increases as the percentage of outliers increases. So the median is preferred. However, there are many other values of  $p$  that could be chosen. With  $p = 0.50$  producing low values of  $R$ , it was expected that at some  $p$  we would achieve maximum values of  $R$  before again seeing a decrease. Therefore, for one of the  $\kappa$  and  $n$  combinations, a simulation was done with more possible values of  $p$ . Figure 8 shows the relationship between  $V$  and  $R$  for  $p = 0.06, 0.12, 0.18, 0.24, 0.30, 0.36, 0.43,$  and  $0.50$ , where  $\kappa = 50$  and  $n = 100$ . From the plot,



**Figure 8.** Plot of  $R$  against  $V$  for 1000 simulated data sets using  $\kappa = 50$  and  $n = 100$ .

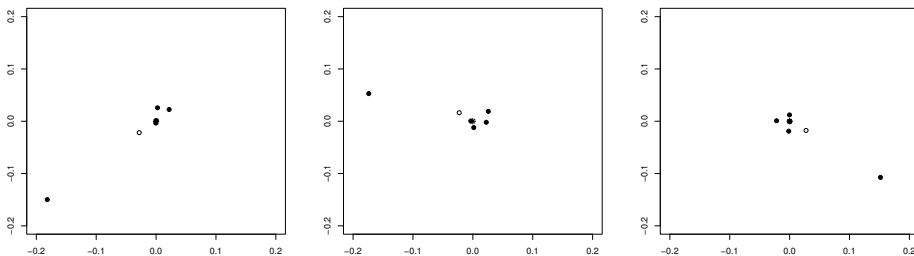


**Figure 9.** Five repeated wrist orientations for the drilling study (represented as a set of three points) and the median rotation (represented as a set of three perpendicular axes).

we see that the median most outperforms the mean near  $p = 0.30$ . As  $p$  increases from 0.30 to 0.50 the value of  $R$  begins to decrease. Even though the median is robust to outlying values, as we let  $p$  approach 0.50, it is ambiguous as to which observations would comprise the “outliers” (with  $p = 0.50$  being a situation that might be best labeled as “bimodal” rather than having outliers at all).

#### 4. Application to drilling data

Now that we have investigated the effectiveness of the median under various parameter choices, we return to the drilling data of [Rancourt et al. 2000]. In Section 1 it was seen that the mean orientation for the five repeated wrist orientations was pulled toward the outlying value. Thus, for this data set, it is desirable to use the median rotation as a measure of center. Figure 9 shows the five repeated wrist orientations on the sphere with the set of perpendicular axes now representing the median rotation. Figure 10 shows the data as a stereographic projection, with the



**Figure 10.** Stereographic projections of the five repeated wrist orientations, in black, with the center point at (0,0) representing the median and the open circle representing the mean.

point in the center at  $(0, 0)$  corresponding to the median rotation and the open circle representing the mean rotation. Both figures show the median near the center of four of the orientations, illustrating the fact that the median is not affected by the outlying value like the mean is.

This small data set with repeated drilling rotations is just one example of a situation in which a median estimator would be preferred over the mean estimator that is typically used to measure the center for three-dimensional rotation data. In subject areas where three-dimensional rotations are common, like the study of human motion, data sets with outliers are bound to show up, making the median estimator developed in Section 2 an important addition to statistics for three-dimensional rotation data.

### References

- [Bingham et al. 2009] M. A. Bingham, D. J. Nordman, and S. B. Vardeman, “Modeling and inference for measured crystal orientations and a tractable class of symmetric distributions for rotations in three dimensions”, *J. Amer. Statist. Assoc.* **104**:488 (2009), 1385–1397. MR 2011a:62189 Zbl 1205.62215
- [Khatri and Mardia 1977] C. G. Khatri and K. V. Mardia, “The von Mises–Fisher matrix distribution in orientation statistics”, *J. Roy. Statist. Soc. Ser. B* **39**:1 (1977), 95–106. MR 58 #13506 Zbl 0356.62044
- [Lee 1995] Y.-S. Lee, “Graphical demonstration of an optimality property of the median”, *The American Statistician* **49**:4 (1995), 369–372.
- [León et al. 2006] C. A. León, J.-C. Massé, and L.-P. Rivest, “A statistical model for random rotations”, *J. Multivariate Anal.* **97**:2 (2006), 412–430. MR 2234030 Zbl 1085.62066
- [Mardia and Jupp 2000] K. V. Mardia and P. E. Jupp, *Directional statistics*, John Wiley & Sons Ltd., Chichester, 2000. MR 2003b:62004 Zbl 0935.62065
- [Morawiec 2004] A. Morawiec, *Orientations and rotations: computations in crystallographic textures*, Springer, Berlin, 2004. MR 2006b:74017 Zbl 1084.74002
- [Rancourt et al. 2000] D. Rancourt, L.-P. Rivest, and J. Asselin, “Using orientation statistics to investigate variations in human kinematics”, *J. Roy. Statist. Soc. Ser. C* **49**:1 (2000), 81–94. MR 1817876 Zbl 0974.62107

Received: 2012-08-03      Revised: 2013-08-30      Accepted: 2013-09-13

mbingham@uwlax.edu      *Department of Mathematics, University of Wisconsin - La Crosse, 1725 State Street, La Crosse, WI 54601, United States*

zach.fischer@milliman.com      *Milliman, 15800 W. Bluemound Road, Suite 100, Brookfield, WI 53005, United States*

# Numerical results on existence and stability of steady state solutions for the reaction-diffusion and Klein–Gordon equations

Miles Aron, Peter Bowers, Nicole Byer, Robert Decker,  
 Aslihan Demirkaya and Jun Hwan Ryu

(Communicated by John Baxley)

In this paper, we study numerically the existence and stability of the steady state solutions of the reaction-diffusion equation,  $u_t - au_{xx} - u + u^3 = 0$ , and the Klein–Gordon equation,  $u_{tt} + cu_t - au_{xx} - u + u^3 = 0$ , with the boundary conditions:  $u(-1) = u(1) = 0$ . We show that as  $a$  varies, the number of steady state solutions and their stability change.

## 1. Introduction

Reaction-diffusion systems are mathematical models which describe the density/concentration of a substance, a population, etc. The typical form is

$$u_t = a\Delta u + f(u), \quad (1)$$

where  $u(x, t)$  is the state variable at position  $x$  and at time  $t$ .  $\Delta u$  is the diffusion term with the diffusion constant  $a$ , and  $f(u)$  is the reaction term.

To motivate Equation (1), consider letting  $a = 0$ , and  $f(u) = ku$ , and we get perhaps the simplest population growth ordinary differential equation  $u_t = ku$ . Here  $u$  represents the size of a population at time  $t$  which is growing with instantaneous growth rate  $k$ . Now imagine taking a one-dimensional spatially distributed population (such as fish in a long, narrow river) and sectioning it into  $N$  subpopulations lined up along the length of the river, each obeying  $u_t = ku$ . In this case, the fish from one subpopulation cannot move to an adjacent subpopulation. When we add the assumption that the fish can move between adjacent subpopulations and in fact will tend to move from more dense to less dense neighboring subpopulations (as is the case with diffusion), and taking the limit as  $N \rightarrow \infty$ , a one-dimensional linear reaction-diffusion equation  $u_t = ku + au_{xx}$  is obtained. The  $au_{xx}$  term has the

*MSC2010:* 35B30, 35B32, 35B35, 35K57, 35L71.

*Keywords:* reaction-diffusion, Klein–Gordon equation, stability, steady state solutions.

effect of limiting growth at point  $x$  if  $u(x, t)$  is concave down as a function of  $x$  (for fixed  $t$ ) at that point ( $a$  assumed positive). Thus for example, if the neighbors of a subpopulation have smaller population densities, that subpopulation will tend to grow smaller due to emigration.

Thus we take the point of view that we can start with a standard dynamical system represented by a first-order differential equation and convert it into a distributed system represented by a partial differential equation by adding a diffusion term  $au_{xx}$ . The situation is similar with dynamical systems represented by second-order differential equations. For example, the equation for a damped mass-spring equation (with no driving force) is well known to be  $mu_{tt} + cu_t + ku = 0$ , where  $u$  represents the displacement from rest of a mass attached to a fixed point by a spring and damper as a function of time  $t$ , and  $m$ ,  $c$ , and  $k$  are parameters representing the mass, the damping constant and the spring constant, respectively. By adding a diffusion term, we get the standard linear Klein–Gordon equation  $mu_{tt} + cu_t + ku = au_{xx}$ , which one can imagine to be a series of (vertical) harmonic oscillators tied together (horizontally) by more springs.

In this paper we are concerned with nonlinear systems, so in the case of the reaction-diffusion equation, instead of starting with the linear differential equation  $u_t = ku$  we start with the nonlinear one:  $u_t = u - u^3$  (which has stable fixed points at  $u = \pm 1$  and an unstable fixed point at  $u = 0$ ). This could be a model of a population with two stable values (after rescaling). This is similar to the example based on the classic spruce-budworm model studied in [Khain et al. 2010]. Converting this ODE to a PDE with the recipe of adding a diffusion term, we get the reaction-diffusion equation known as the Allen–Cahn equation:

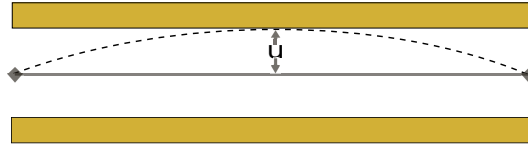
$$u_t = au_{xx} + u - u^3. \quad (2)$$

For the case of the mass-spring system, we replace the linear restoring force by a nonlinear force of the form  $f(u) = \alpha u + \beta u^3$  and end up with the Duffing equation:

$$u_{tt} + cu_t + \alpha u + \beta u^3 = 0, \quad (3)$$

where  $c$  is damping, and  $\alpha$  and  $\beta$  are chosen so as to model various physical systems. For example, if  $\alpha$  is negative and  $\beta$  is positive, the force tends to move the mass away from  $u = 0$  and towards  $u = \pm 1$  (one system with this property would be when magnets are added above and below the mass, which is assumed to be, say, made of iron, so that the mass is pulled from its equilibrium position either up or down, see Figure 1). Again, adding a diffusion term we get a nonlinear Klein–Gordon equation:

$$u_{tt} + cu_t + \alpha u + \beta u^3 = au_{xx}, u(-1) = u(1) = 0. \quad (4)$$



**Figure 1.** Damped, thin metal wire, fixed at  $x = \pm 1$ , placed between two magnets at  $u = \pm 1$ .

In both equations above when the parameter  $a$  is very small, we are back to a bunch of unconnected one-dimensional systems behaving independently. For the fish population in a river model, this would represent the idea that the fish could not migrate up or down the river, but can only reproduce and/or die in one location. For the mass-spring system this would represent totally unconnected, side-by-side oscillators with magnets. Thus we refer to  $a$  as the *strength of connection* parameter.

In this paper, by using *spectral methods*, we study the numerical existence and the stability of the steady state solutions of reaction-diffusion equation:

$$u_t = au_{xx} + u - u^3 \tag{5}$$

and Klein–Gordon equation:

$$u_{tt} + cu_t = au_{xx} + u - u^3. \tag{6}$$

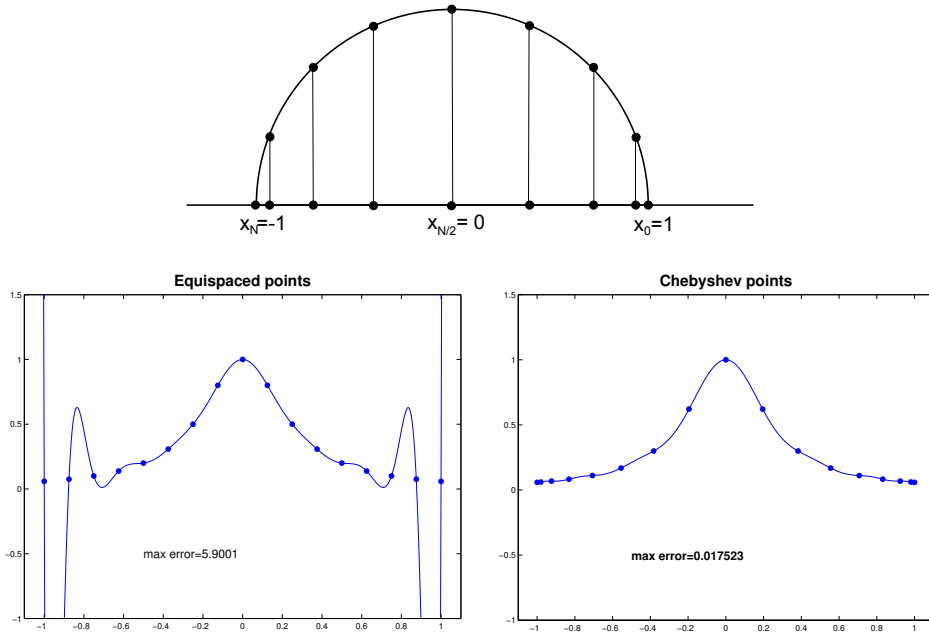
For both cases we take the boundary conditions as  $u(-1) = u(1) = 0$ . We show that as  $a$  varies, the number of those special solutions and their stability change.

## 2. Spectral methods

**Introduction.** Because of the nonlinear terms added to our second order partial differential equations, we chose to use a numerical method of analysis instead of finding exact solutions. We used spectral methods of analysis (see [Trefethen 2000]) instead of more traditional methods, such as finite differences, due to the exponential order of error convergence that the spectral methods demonstrate.

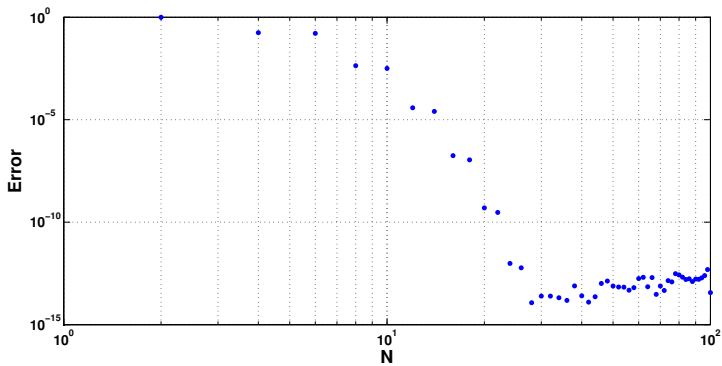
Spectral methods break the second order partial differential equation into a series of first order differential equations. Each first order differential equation lies at a point called a Chebyshev point, which is similar to the equally spaced points used in finite differences. However Chebyshev points are selected by taking the  $x$ -coordinates of equally spaced points on a half-circle Figure 2, top.

Chebyshev points are closer together towards the endpoints of the equation, which provides a much better polynomial fit and therefore much greater accuracy compared to equally spaced points as seen in Figure 2, bottom (reproduced from [Trefethen 2000]).



**Figure 2.** Top: Chebyshev points. Bottom: Chebyshev versus equispaced points.

Spectral methods have an exponential order of error convergence, so the error decreases much more rapidly than in other numerical methods. By increasing  $N$ , defined as the number of Chebyshev points, linearly, the error converges exponentially as in Figure 3 (reproduced from [Trefethen 2000]). However with finite differences method, the number of points are normally increased by a factor of ten to gain just one more decimal point of accuracy.



**Figure 3.** Convergence of spectral differentiation.



**Choosing  $N$ .** In order to find the most accurate solution, picking the correct number of Chebyshev points is imperative. As discussed earlier, as  $N$  is increased linearly the error decreases exponentially but only to a certain point. Eventually the error reaches a minimum and then begins to increase slowly due to machine error.

**Claim 2.1.** *For the reaction diffusion equation (5), the minimum error occurs at approximately  $N = 14$ .*

*Proof.* In order to select the optimal number of Chebyshev points, we calculate the error by comparing the exact solution of the linear form of (5) to the solution calculated with spectral methods. The linear form of (5) is

$$u_t = au_{xx} + u, \quad u(-1) = u(1) = 0. \tag{7}$$

By using the *separation of variables method*, we get the general solution of (7) as

$$u(x, t) = \sum_{n=1}^{\infty} c_n e^{(1-a/4(n\pi)^2)t} \sin(n\pi(x+1)/2). \tag{8}$$

If we pick our initial data as the eigenfunction  $u(x, 0) = \sin(\pi(x+1)/2)$ , we get  $c_1 = 1$  and  $c_n = 0$  for  $n \geq 2$ . So the exact solution of (7) with that initial data is

$$u(x, t) = e^{(1-a/4\pi^2)t} \sin(\pi(x+1)/2). \tag{9}$$

The approximate solution of (7) derived from the spectral methods:

$$u(x, t) = u_0 e^{At},$$

where  $A = aD^2 + 1$ ,  $u_0 = \sin \pi(x+1)/2$  and  $D^2$  is the Chebyshev matrix derived by spectral methods.

Now we define the error as

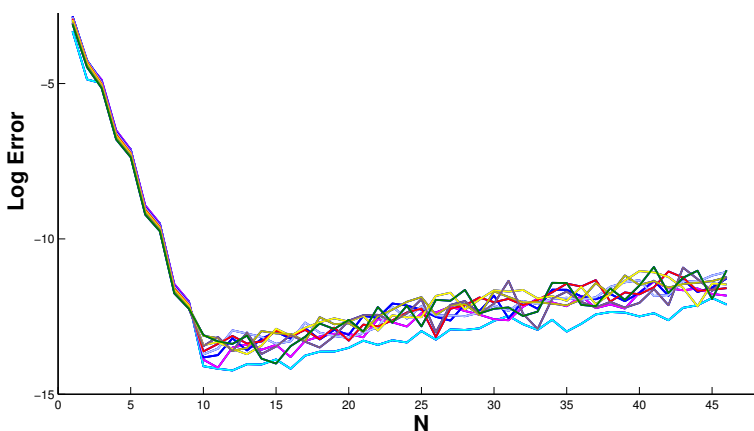
$$\begin{aligned} \text{Error} &= \|\text{Exact solution} - \text{Approximate solution}\| \\ &= \left\| \left( e^{(1-a/4\pi^2)t} - e^{(aD^2+1)t} \right) \sin(\pi(x+1)/2) \right\|. \end{aligned}$$

where  $\|\cdot\|$  represents the Euclidean norm. In Figure 4, for various  $a$  values, we observe that the error reaches its minimum at approximately  $N = 14$ . □

**Claim 2.2.** *For the Klein-Gordon equation (6), the minimum error occurs at approximately  $N = 12$ .*

*Proof.* Similar to what we did for the reaction-diffusion equation, we calculate the error by comparing the exact solution of the linear form of (6) to the solution calculated with spectral methods. The linear form of (6) is

$$u_{tt} + cu_t = au_{xx} + u, \quad u(-1) = u(1) = 0. \tag{10}$$



**Figure 4.** Error plots for  $a$ -values from 0.02 to 0.4.

By using the *separation of variables method* and picking the initial data as

$$u(x, 0) = \sin \frac{\pi(x + 1)}{2} \tag{11}$$

and

$$u_t(x, 0) = \left( -\frac{c}{2} + \sqrt{\frac{c^2}{4} + 1 - \frac{a}{4} \pi^2} \right) \sin \frac{\pi(x + 1)}{2} \tag{12}$$

and assuming  $c^2 + 4 - a\pi^2 > 0$ , we get the exact solution of (10) as

$$u(x, t) = \sin \left( \frac{\pi(x + 1)}{2} \right) \exp \left( \left( -\frac{c}{2} + \sqrt{\frac{c^2}{4} + 1 - \frac{a}{4} \pi^2} \right) t \right). \tag{13}$$

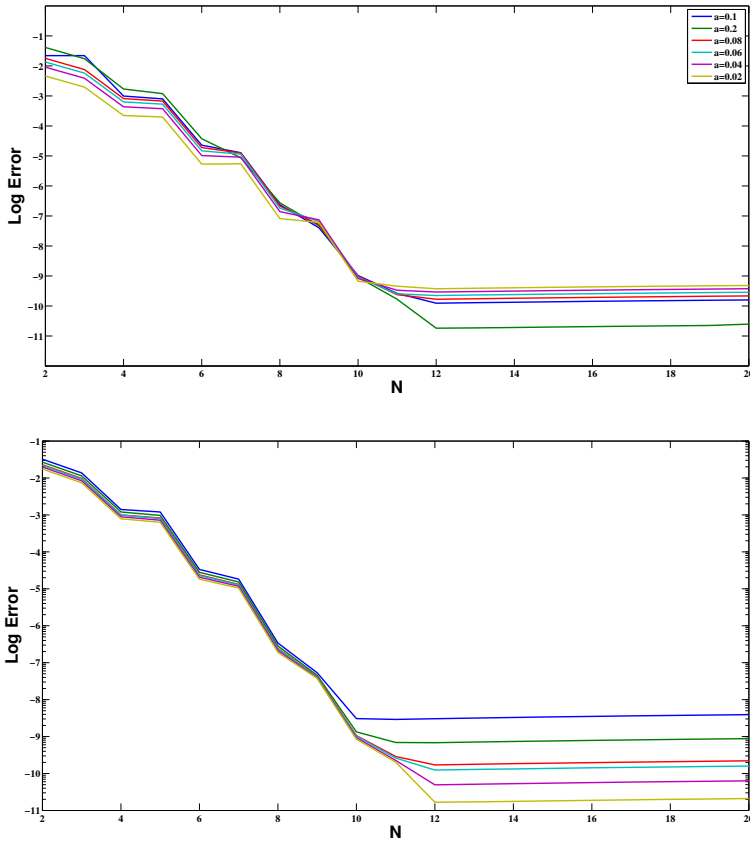
The approximate solution of (10) derived from the spectral methods is found by using ODE45 by changing the second order differential equation

$$u_{tt} + cu_t = au_{xx} + u = (aD^2 + 1)u$$

into a first order system  $y$ , defining  $z := u_t$ ,

$$\begin{bmatrix} u \\ z \end{bmatrix}_t = \begin{bmatrix} 0 & I \\ aD^2 + I & -c \end{bmatrix} \begin{bmatrix} u \\ z \end{bmatrix}$$

and using the same initial conditions (11) and (12). We define the error same as we defined for the reaction-diffusion equation and use the Euclidean norm. Figure 5 shows the error for various  $a$  values and fixed  $c = 1$  (top), an for various  $c$  values and fixed  $a = 0.1$  (bottom); we observe that the error reaches its minimum at approximately  $N = 12$ . □



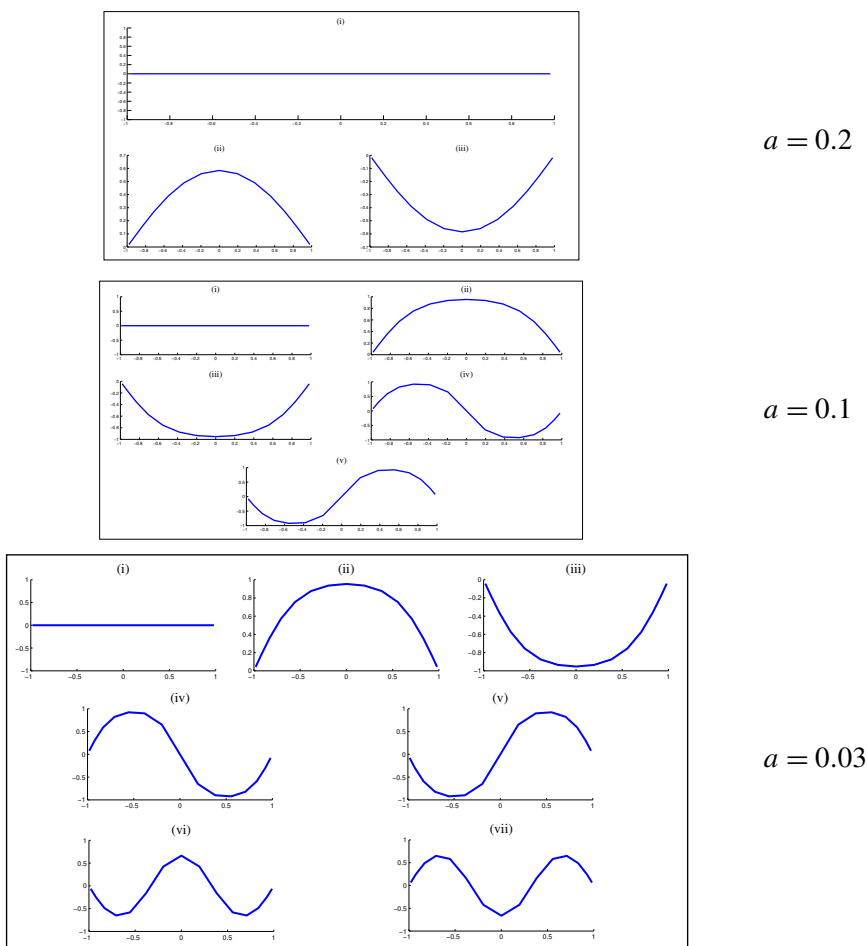
**Figure 5.** Log error. Top:  $c = 1$ , values of  $a$  from 0.02 to 0.2. Bottom:  $a = 0.1$ , values of  $c$  from 0.1 to 2.

### 3. Numerical existence of steady state solutions

**Steady state solutions.** For both equations, the reaction-diffusion (5) and the Klein-Gordon equation (6), the steady state solutions  $u(x, t) = \phi(x)$  satisfy the same equation

$$-a\phi'' - \phi + \phi^3 = 0, \quad \phi(-1) = \phi(1) = 0 \tag{14}$$

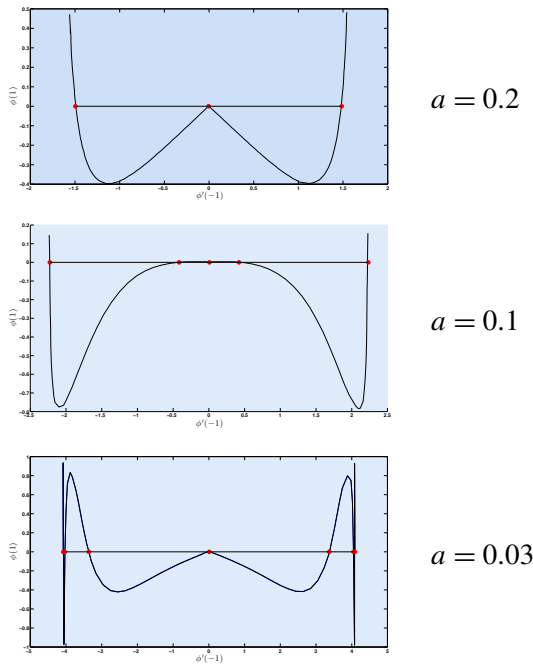
since  $\phi_t = 0$  and  $\phi_{tt} = 0$ . Figure 6 shows the steady state solutions for three different  $a$ -values,  $a = 0.2$ ,  $a = 0.1$  and  $a = 0.03$ . Consecutively (14) has 3, 5 and 7 solutions. The numerical computations show that as  $a$  is decreased, two new steady states of opposite sign and increasing number of oscillations occur for each bifurcation. The relation between  $a$  and number of steady state solutions will be analytically studied on page 731.



**Figure 6.** Steady state solutions for various values of  $a$ .

Note that zero is the trivial steady state solution for any  $a$ . For nonzero solutions, throughout this paper, we will name the steady state solutions. For example, we will name each convex steady state solution in Figure 6 an “n” solution — see graphs labeled (ii); each concave solution a “u” solution — graphs labeled (iii); graphs labeled (iv) for  $a = 0.1$  and  $a = 0.03$  are the “nu” solutions, and so on.

**Bifurcations.** The number of steady states for a given  $a$  value was verified with the shooting method. In Figure 7 we see shooting method plots for three  $a$  values. The number of solutions at an  $a$  value is the number of times the plot of  $\phi(1)$  versus  $\phi'(-1)$  touches the  $\phi'(-1)$  axis where  $\phi(1) = 0$ . The bifurcation values to five decimal places were determined by changing  $a$  until a new number of solutions was observed. These bifurcation values are confirmed in the next subsection.



**Figure 7.** Existence of steady states  $\phi(1)$  vs  $\phi'(-1)$ , for various values of  $a$ .

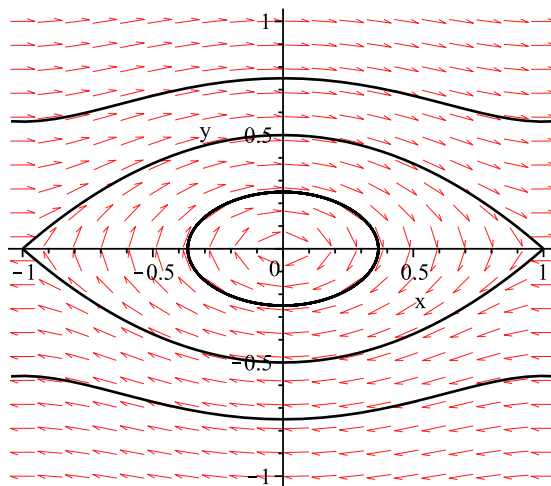
**Determination of bifurcation values.** In Section 3 and in the preceding subsection, the steady state solutions to both the nonlinear reaction diffusion equation (5) and the nonlinear Klein–Gordon equation (6) were calculated numerically, and the number of solutions were found for three  $a$  values. In particular it was found that for  $a = 0.2$  there are three steady state solutions, for  $a = 0.1$  there are five steady state solutions and for  $a = 0.03$  there are seven steady state solutions.

In this section we find all bifurcation values for the parameter  $a$ ; at each value two new solutions are added. We will show that these bifurcations values are at  $a = (2/n\pi)^2$  for  $n$  a positive integer. Thus the first few values are 0.4053, 0.1013, 0.0450, 0.0253; for  $a > 0.4503$  there is only the zero solution, for  $0.1013 < a < 0.4503$  there are three solutions, for  $0.0450 < 0.1013$  there are five solutions, for  $0.0253 < a < 0.0450$  there are seven solutions, and so on. This is consistent with the numerical results.

Consider the initial value problem

$$x' = y, \quad y' = -\lambda^2(x - x^3), \quad x(-1) = 0, \quad y(-1) = y_0, \quad (15)$$

which is equivalent to equations (14) with  $1/\lambda^2$  substituted for  $a$ .



**Figure 8.** Phase portrait for  $x' = y$ ,  $y' = -\lambda^2(x - x^3)$  with  $\lambda = 0.5$ .

Clearly this system has saddle points at  $(\pm 1, 0)$ . Due to the symmetry of the vector field for this system, the fixed point at  $(0, 0)$  is a center. Solution curves closer to the origin than the stable manifolds of  $(\pm 1, 0)$  form closed loops. Thus the solution curves to (15) circle the origin in the clockwise direction for  $y_0$  sufficiently small. See Figure 8.

Let  $(x(y_0, t), y(y_0, t))$  represent the solution to (15) and let  $\theta(y_0, t)$  represent the angle that the line segment connecting  $(x(y_0, t), y(y_0, t))$  with  $(0, 0)$  makes with the positive  $x$ -axis. Thus  $\theta(y_0, t)$  is the angle in the polar coordinate representation of  $(x(y_0, t), y(y_0, t))$ , and hence  $\tan \theta(y_0, t) = y(y_0, t)/x(y_0, t)$ . Assume that  $\theta$  starts at  $\pi/2$ , corresponding to the initial condition given in (15), and continues to decrease as the solution curve moves clockwise around the origin. Thus after one loop of the solution curve around the origin  $\theta$  is  $-3\pi/2$ , after two loops  $\theta$  is  $-7\pi/2$ , and so on.

**Theorem 3.1.**  $\theta(y_0, t)$  is an increasing function of  $y_0 > 0$  for fixed  $t$ .

*Proof.* For convenience we suppress the  $y_0$  argument and write  $\theta(y_0, t)$  as  $\theta(t)$ . Differentiating  $\tan \theta(t) = y(t)/x(t)$  with respect to  $t$  we get

$$(1 + \tan^2 \theta(t))\theta'(t) = \frac{y'(t)x(t) - y(t)x'(t)}{x^2(t)}.$$

Solving for  $\theta'(t)$  and using  $\tan^2 \theta(t) = y(t)^2/x(t)^2$  results in

$$\theta'(t) = \frac{y'(t)x(t) - y(t)x'(t)}{x^2(t) + y^2(t)}.$$

Then using the DE system  $x' = y$ ,  $y' = -\lambda^2(x - x^3)$  we get

$$\theta'(t) = \frac{-\lambda^2 x^2(t) + \lambda^2 x^4(t) - y^2(t)}{x^2(t) + y^2(t)}.$$

Switching to polar coordinates ( $x = r \cos \theta$ ,  $y = r \sin \theta$ ,  $r^2 = x^2 + y^2$ ) on the right side yields

$$\begin{aligned} \theta'(t) &= \frac{-\lambda^2 r^2(t) \cos^2 \theta(t) + \lambda^2 r^4(t) \cos^4 \theta(t) - r^2(t) \sin^2 \theta}{r^2(t)} \\ &= -\lambda^2 \cos^2 \theta(t) + \lambda^2 r^2(t) \cos^4 \theta(t) - \sin^2 \theta(t). \end{aligned}$$

Rearranging a bit we get

$$\theta'(t) = \lambda^2 \cos^2 \theta(t) (-1 + r^2(t) \cos^2 \theta(t)) - \sin^2 \theta(t). \quad (16)$$

Using  $x = r \cos \theta$  we could also write (16) as

$$\theta'(t) = \lambda^2 \cos^2 \theta(t) (-1 + x^2(t)) - \sin^2 \theta(t), \quad (17)$$

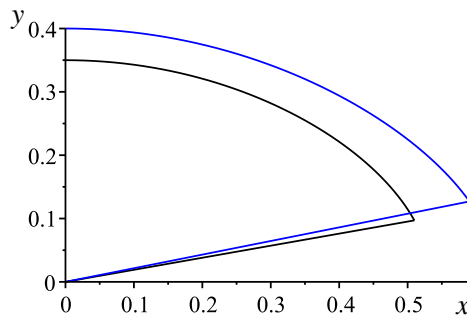
which shows that  $\theta' < 0$  for  $-1 < x < 1$ . This is to be expected as we know that solution curves inside the unstable manifold of the fixed points circle the origin clockwise.

It is clear from (16) that  $\theta'$  increases as a function of  $r$  for fixed  $\theta$ . Since  $\theta'$  is negative this means that for a given  $\theta$ , the solution curves farther from the origin are circling the origin at a slower angular rate (smaller absolute value) than those that are closer. See Figure 9.

This implies that for  $y_0$  chosen so that the solution curve forms a closed loop, smaller  $y_0$  means that the solution curve has wrapped further around the origin in the clockwise direction, and hence  $\theta(y_0, t)$  is smaller (for fixed  $t$ ). This means  $\theta(y_0, t)$  is an increasing function of  $y_0$  as claimed.

Let  $\lambda > 0$  be fixed. Let  $x_1(y_0, t)$  represent the solution to

$$x' = y, \quad y' = -\lambda^2 x, \quad x(-1) = 0, \quad y(-1) = y_0. \quad (18)$$



**Figure 9.** Solution curves closer to the origin move faster.

This is just (15) without the  $x^3$  term. It is easy to show that

$$x_1(y_0, t) = y_0 \sin(\lambda(t+1))/\lambda \quad \text{and} \quad y_1(y_0, t) = y_0 \cos(\lambda(t+1)).$$

As before, let  $(r_1, \theta_1)$  be the polar representation of  $(x_1, y_1)$ , so that  $\theta_1(y_0, t)$  is the polar angle for the point  $(x_1(y_0, t), y_1(y_0, t))$ . Since

$$\tan \theta_1(y_0, t) = \frac{y_1(y_0, t)}{x_1(y_0, t)} = \frac{y_0 \cos(\lambda(t+1))}{y_0 \sin(\lambda(t+1))/\lambda} = \lambda \cot(\lambda(t+1)),$$

we know that  $\theta_1(y_0, t)$  is in fact independent of  $y_0$ . Thus instead of  $\theta_1(y_0, t)$  we will write  $\theta_1(t)$ .  $\square$

**Theorem 3.2.** Fix  $t$  and  $\lambda$ . As  $y_0 \rightarrow 0$ ,  $\theta(y_0, t) \rightarrow \theta_1(t)$  monotonically.

*Proof.* The monotonic part follows from the previous theorem. To prove the limit part, the basic idea is that  $x^3$  is negligible compared to  $x$  for small  $x$ , and so the linear and nonlinear vector fields, given in Equations (15) and (18), respectively, are indistinguishable for small  $x$ . We now flesh out the details of this argument.

For the IVP,  $x'' + \lambda^2(x - x^3) = 0$ ,  $x(-1) = 0$ ,  $x'(-1) = y_0$ , which is just (15) written in second-order form, we can multiply the DE by  $x'$  to get

$$x'x'' + \lambda^2(x - x^3)x' = 0.$$

We can integrate both sides now to get

$$\frac{1}{2}(x')^2 + \lambda^2\left(\frac{1}{2}x^2 - \frac{1}{4}x^4\right) = C.$$

Then using  $x(-1) = 0$  and  $x'(-1) = y_0$  we get  $C = y_0^2/2$ . We now have

$$(x')^2 + \lambda^2\left(x^2 - \frac{1}{2}x^4\right) = y_0^2 \tag{19}$$

after substituting and multiplying by 2. Plotting (19) in the phase plane ( $x$  on the horizontal axis and  $x'$  on the vertical) for various  $y_0$  we are back to the closed curves in Figure 8, which shows clearly that  $x(t)$  can be “trapped” in an arbitrarily small region  $-\varepsilon < x(t) < \varepsilon$  if  $y_0$  is taken to be sufficiently small.  $\square$

In order to finish the proof of the theorem we need to invoke a version of the Gronwall inequality:

**Lemma 3.3** (Gronwall’s inequality: see for example [Howard 1998, Theorem 2.1]). Let  $X$  be a Banach space and  $U \subset X$  an open set in  $X$ . Let  $f, g : [a, b] \times U \rightarrow X$  be continuous functions and let  $y, z : [a, b] \rightarrow U$  satisfy the initial value problems

$$\begin{aligned} y'(t) &= f(t, y(t)), \quad y(a) = y_0 \\ z'(t) &= g(t, z(t)), \quad z(a) = z_0. \end{aligned} \tag{20}$$



Also assume there is a constant  $C \geq 0$  so that

$$\|g(t, x_2) - g(t, x_1)\| \leq C \|x_2 - x_1\| \tag{21}$$

and a continuous function  $\phi : [a, b] \rightarrow [0, \infty)$  so that

$$\|f(t, y(t)) - g(t, y(t))\| \leq \phi(t). \tag{22}$$

Then for  $t \in [a, b]$

$$\|y(t) - z(t)\| \leq e^{C|t-a|} \|y_0 - z_0\| + e^{C|t-a|} \int_a^t e^{-C|s-a|} \phi(s) ds.$$

For our purposes the Banach space  $X$  is just the real numbers, so that the norm is the absolute value.

For convenience we suppress the  $y_0$  and write  $\theta(t)$  instead of  $\theta(y_0, t)$ . We can rewrite (17) as

$$\theta'(t) = f(t, \theta(t)),$$

where

$$f(t, \theta) = -\lambda^2 \cos^2 \theta - \sin^2 \theta + \lambda^2 x^2(t) \cos^2 \theta$$

and where  $x(t)$  comes from the solution to the full system in (15). Similarly,  $\theta_1$  satisfies

$$\theta_1'(t) = g(t, \theta_1(t)),$$

where

$$g(t, \theta) = -\lambda^2 \cos^2 \theta - \sin^2 \theta.$$

We now apply the Gronwall inequality using the above choices for  $f$  and  $g$ , where the interval  $[a, b]$  is  $[-1, 1]$  and where we choose the same initial condition  $\pi/2$  for the DEs, that is,  $\theta(-1) = \pi/2$  and  $\theta_1(-1) = \pi/2$  (which corresponds to  $x(0) = 0$  and  $y(0) = y_0$  in rectangular coordinates). Equation (21) is called a Lipschitz condition and is satisfied by  $g(t, \theta)$  for some  $C$  because it is continuously differentiable as a function of  $\theta$ . Finally since

$$|f(t, \theta(t)) - g(t, \theta(t))| = |\lambda^2 x^2(t) \cos^2 \theta(t)| \leq \lambda^2 x^2(t),$$

we see that (22) is satisfied with  $\phi(t) = \lambda^2 x^2(t)$ . The conclusion of the Gronwall inequality follows, which means that for fixed  $t \in [-1, 1]$  and  $\lambda$

$$|\theta(t) - \theta_1(t)| \leq e^{C|t+1|} \int_{-1}^t e^{-C|s+1|} \lambda^2 x^2(s) ds$$

since the initial conditions are the same. If we assume that  $|x(t)| < \varepsilon$  then

$$\begin{aligned} e^{C|t+1|} \int_{-1}^t e^{-C|s+1|} \lambda^2 x^2(s) ds &< e^{C|t+1|} \int_{-1}^t e^{-C|s+1|} \lambda^2 \varepsilon^2 ds \\ &= \frac{1}{C} \lambda^2 \varepsilon^2 (e^{C|t+1|} - 1). \end{aligned}$$

Following the above inequalities we have shown that

$$|\theta(y_0, t) - \theta_1(t)| < \frac{1}{C} \lambda^2 \varepsilon^2 (e^{C|t+1|} - 1).$$

The theorem follows by recalling that  $|x(t)|$  can be made arbitrarily small for  $y_0$  sufficiently small.

**Theorem 3.4.** *The eigenvalues of the linear BVP*

$$x'' + \lambda^2 x = 0, \quad x(-1) = 0, \quad x(1) = 0 \quad (23)$$

*are the bifurcation points for the nonlinear BVP*

$$x'' + \lambda^2(x - x^3) = 0, \quad x(-1) = 0, \quad x(1) = 0. \quad (24)$$

*Specifically the eigenvalues of the linear problem are  $\lambda_n = n\pi/2$ , and there is one solution to the nonlinear problem (the zero solution) for  $0 \leq \lambda \leq \lambda_1$ , three solutions to the nonlinear problem for  $\lambda_1 < \lambda \leq \lambda_2$ , five solutions to the nonlinear problem for  $\lambda_2 < \lambda \leq \lambda_3$ , and so on.*

*Proof.* If  $\lambda_n$  is an eigenvalue for the BVP in (23) it is easy to show that  $\lambda_n = n\pi/2$  for  $n = 1, 2, \dots$  and that the corresponding eigenfunctions are any multiple of  $\sin(\lambda_n(t+1))$ . Note that the system of differential equations in (18) is equivalent to the differential equation in (23). Thus if  $\lambda = \lambda_n$  for some  $n \geq 1$ , then a solution to (18) automatically satisfies  $x(1) = 0$  and so is a solution to (23) for any  $y_0$ . This explains the relationship between the IVP in (18) and the BVP in (23).

The relationship between the nonlinear IVP in (15) and the nonlinear BVP in (24) is similar in that the differential equations are equivalent. However, because of the nonlinearity, (24) can be solved for *any*  $\lambda$  by solving the IVP in (15) and varying  $y_0$  until  $x(1) = 0$  is obtained (sometimes called the “shooting method” for solving a BVP). Note that  $x(1) = 0$  in the phase plane means that the angle  $\theta(y_0, 1)$  is any of  $-\pi/2, -3\pi/2, -5\pi/2, \dots$

Now fix  $\lambda$  and start with  $y_0$  chosen so that the solution to the IVP in (15) is the stable manifold of  $(1, 0)$ . As  $y_0$  decreases towards zero,  $\theta(y_0, 1)$  will wrap clockwise around the origin until it reaches  $\theta_1(1)$  in the limit, as shown in the previous theorem. Every time  $\theta(y_0, 1)$  passes  $-n\pi/2$  for  $n$  odd we get a solution to the nonlinear BVP in (24). If  $\lambda < \lambda_1$  this never happens (because  $\theta_1(1) > -\pi/2$ ), so the only solution is the zero solution. If  $\lambda_1 < \lambda < \lambda_2$  then  $-3\pi/2 < \theta_1(y_0, 1) < -\pi/2$

and so  $\theta(y_0, 1)$  will pass just  $-\pi/2$ , in which case we have the zero solution as well as one more solution. By symmetry, a third solution occurs for  $y_0$  negative. If  $\lambda_2 < \lambda < \lambda_3$ , then  $\theta(y_0, 1)$  will pass  $-\pi/2$  and  $-3\pi/2$  yielding two solutions, plus the zero solution, plus the symmetric solutions for  $y_0$  negative, for a total of five. Proceeding in this manner, the theorem is proved.  $\square$

As stated earlier in this section, we have substituted  $1/\lambda^2$  for the parameter  $a$  in the nonlinear reaction-diffusion and Klein-Gordon equations, and thus the bifurcation values in terms of  $a$  are  $a = (2/n\pi)^2$ .

#### 4. Stability analysis

In order to determine the stability of a steady state solution, we look at the solution in the vicinity of the steady state and observe its behavior over time.

We assume that  $u(x, t)$  is the solution such that  $u_0 = u(x, 0)$  is in the vicinity of the steady state solution. The difference between the solution and steady state solution is known as the perturbation and is denoted as  $v(x, t)$  and, satisfies

$$v(x, t) = u(x, t) - \phi(x). \tag{25}$$

Stability analysis will require us to study the long time behavior of the perturbation. If the solution diverges from the steady state ( $\lim_{t \rightarrow \infty} \|v\| \rightarrow \infty$ ), then the steady state is unstable. If the solution does not diverge and the perturbation remains small, the steady state is stable.

**Stability of reaction-diffusion equation.** By substituting (25) into (5), we get

$$(\phi + v)_t - a(\phi + v)_{xx} - (\phi + v) + (\phi + v)^3 = 0.$$

Since  $-a\phi'' - \phi + \phi^3 = 0$  and  $\phi_t = 0$ , we get

$$v_t - av_{xx} - v + 3\phi^2v + 3\phi v^2 + v^3 = 0.$$

By stability manifold theorem, we can say that the stability of the above equation will be similar to its linearized equation which is as follows:

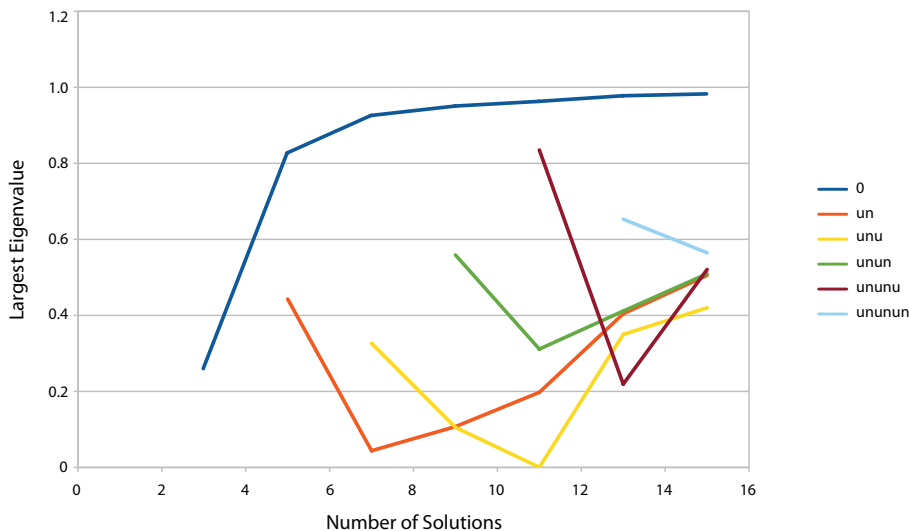
$$v_t - av_{xx} - v + 3\phi^2v = 0,$$

which can be rewritten as

$$v_t = (aD^2 + 1 - 3\phi^2)v = 0.$$

So the eigenvalues of the linearized RD operator  $aD^2 + 1 - 3\phi^2$  will tell us if  $v$  blows up in time, decreases to 0, or stays bounded.

Our numerical results show that the largest eigenvalues of the linearized RD operator about the “u” and “n” solutions are negative. This implies that these



**Figure 10.** Largest eigenvalue of the linearized RD operator for the unstable steady states.

solutions are stable. All other solutions are unstable because the linearized RD operator has at least one positive eigenvalue. Figure 10 shows the largest eigenvalue of the linearized RD operator about each unstable steady state as the number of solutions from a given  $a$ -value increases (as the  $a$ -value decreases).

**Stability of Klein–Gordon equation.** By substituting (25) into (6), we get

$$(\phi + v)_{tt} + c(\phi + v)_t - a(\phi + v)_{xx} - (\phi + v) + (\phi + v)^3 = 0.$$

Since  $-a\phi'' - \phi + \phi^3 = 0$ , and  $\phi_t = 0$  and  $\phi_{tt} = 0$ , we get

$$v_{tt} - av_{xx} - v + 3\phi^2v + 3\phi v^2 + v^3 = 0.$$

By the stable manifold theorem, we can say that the stability of the above equation will be similar to its linearized equation, which is as follows:

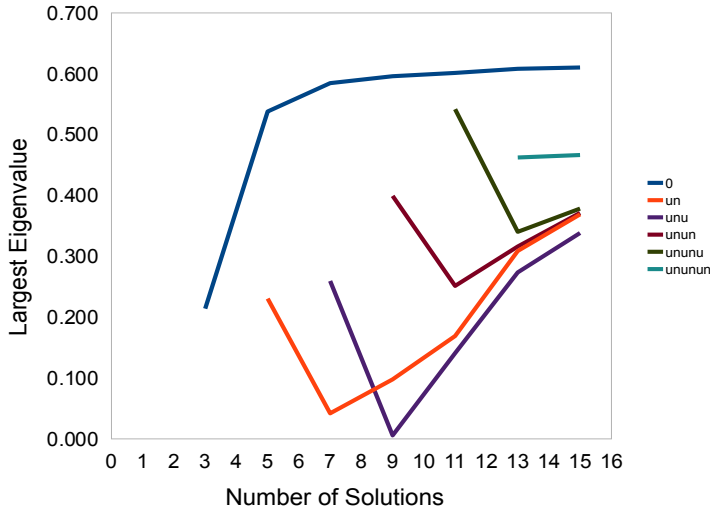
$$v_{tt} + cv_t - av_{xx} - v + 3\phi^2v = 0.$$

Let’s write it as a first order system by defining  $v_t = w$ . We get

$$\begin{bmatrix} v \\ w \end{bmatrix}_t = \begin{bmatrix} 0 & I \\ aD^2 + I - 3\phi^2 & -c \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix}.$$

So we need to find the eigenvalues of the operator matrix in this equation in order to find the long time behavior of  $\begin{bmatrix} v \\ w \end{bmatrix}$ .

Because all of the new steady states that occur after three solutions are unstable, we then look at how unstable they are. One metric for instability is the magnitude



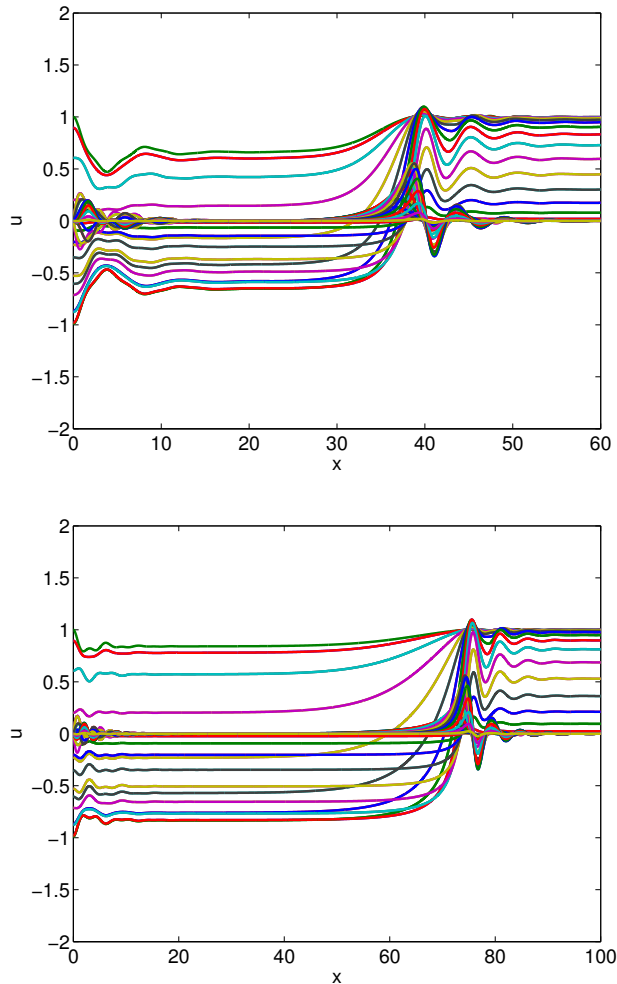
**Figure 11.** Largest eigenvalue of the linearized KG operator for the unstable steady states.

of the largest positive eigenvalue. The larger this eigenvalue is, the more unstable the steady state. Figure 11 shows the magnitude of the largest positive eigenvalue for various solutions as the number of solutions from a given  $a$ -value increases (as the  $a$ -value decreases). Note how the solution becomes more stable at first, and then becomes less stable as  $a$  decreases. Also note that the zero solution becomes more unstable quickly but then approaches a level of instability asymptotically.

In Figure 12, steady states are indicated stable or unstable and are organized by the bifurcation range that they occur in (number of solutions for a range of values

		Number of Solutions								
		1	3	5	7	9	11	13	15	
0	stable	stable	unstable	unstable	unstable	unstable	unstable	unstable	unstable	
			stable	stable	stable	stable	stable	stable	stable	
			stable	stable	stable	stable	stable	stable	stable	
				unstable	unstable	unstable	unstable	unstable	unstable	
				unstable	unstable	unstable	unstable	unstable	unstable	
					unstable	unstable	unstable	unstable	unstable	
						unstable	unstable	unstable	unstable	
							unstable	unstable	unstable	
								unstable	unstable	
									unstable	
										unstable

**Figure 12.** The stability of different types of solutions as the value of  $a$  decreases.



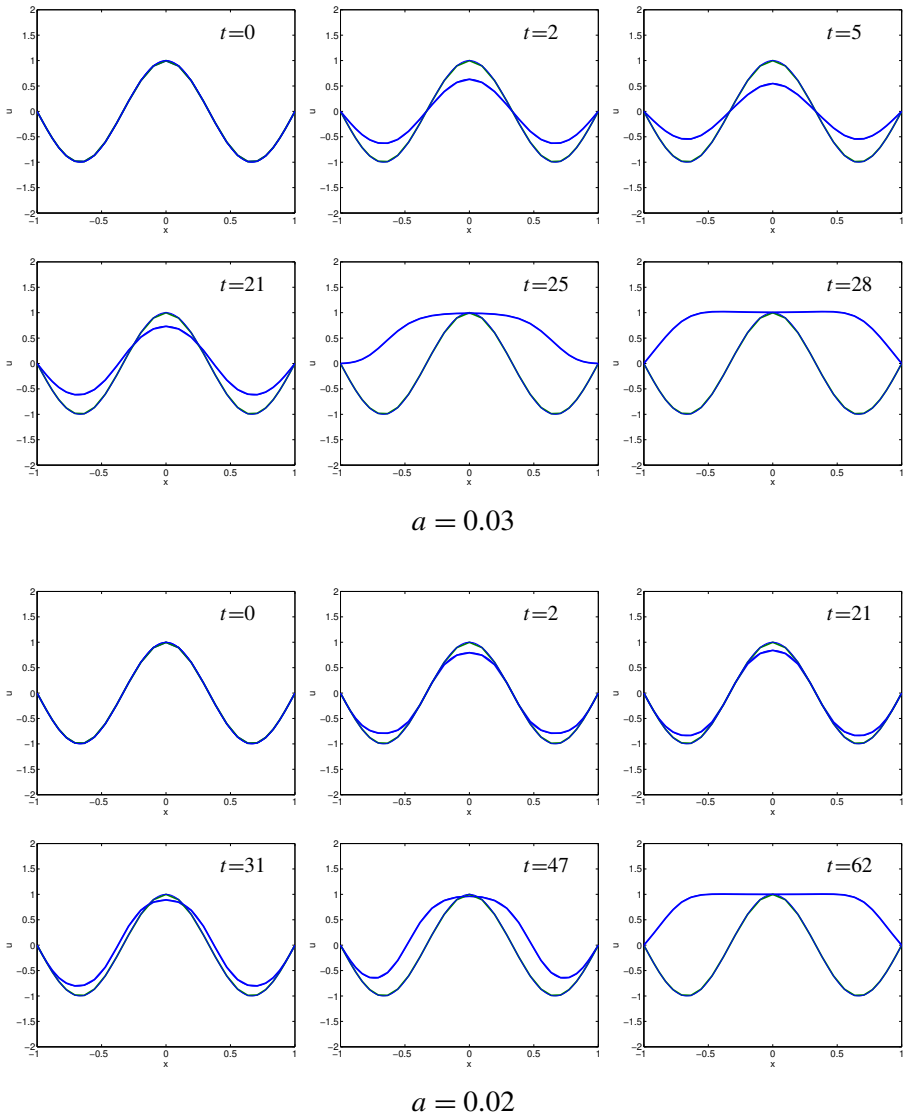
**Figure 13.** Side-view plots. Top:  $a = 0.03$ ,  $c = 0.5$ . Bottom:  $a = 0.02$ ,  $c = 0.5$ .

of  $a$ ) and the look of the solution. For  $a$  values with only one steady state, the only steady state is the zero solution and it is stable. For all other bifurcation ranges, the zero solution is unstable. The solutions depicted in the second and third row of Figure 12 are stable. All other steady states are unstable.

In Figure 13 we compare two side-view plots to verify that the increase in stability of unstable steady states as observed in Figure 11 actually happens. As expected, the side plot with  $a = 0.02$  (9-solution range) took longer to move to a stable solution than the plot with  $a = 0.03$  (7-solution range) because it was less unstable (the positive eigenvalues were closer to being negative).

### 5. Simulations

Figure 14 depicts simulations of the Klein–Gordon equation for the comparison made in the previous section between  $a = 0.03$  and  $a = 0.02$  with  $c = 0.5$ . The simulation with  $a = 0.02$  takes longer than the simulation with  $a = 0.03$  to reach a stable steady state solution. These images agree with the side-view plots in Figure 13 and the plot in Figure 11. The simulations can be seen at [youtu.be/dINbTOUUMX8](https://youtu.be/dINbTOUUMX8) ( $a = 0.02$ ) and [youtu.be/ccdF6tU2Vcw](https://youtu.be/ccdF6tU2Vcw) ( $a = 0.03$ ).



**Figure 14.** Results of the simulations for  $a = 0.03$  and  $a = 0.02$ .

## References

- [Howard 1998] R. Howard, “The Gronwall inequality”, lecture notes, 1998, available at <http://www.math.sc.edu/~howard/Notes/gronwall.pdf>.
- [Khain et al. 2010] E. Khain, Y. T. Lin, and L. M. Sander, “Fluctuations and stability in front propagation”, preprint, 2010. arXiv 1009.5945v1
- [Trefethen 2000] L. N. Trefethen, *Spectral methods in MATLAB*, Software, Environments, and Tools **10**, Society for Industrial and Applied Mathematics, Philadelphia, 2000. MR 2001c:65001 Zbl 0953.68643

Received: 2012-12-05    Revised: 2013-10-29    Accepted: 2013-11-05

Miles.Aron@uzh.ch	<i>University of Zurich, Sonneggstrasse 23, CH-8006 Zurich, Switzerland</i>
pbowers@fas.harvard.edu	<i>Harvard University, 230 Chestnut Street, Cambridge, MA 02139, United States</i>
nicole_byer@brown.edu	<i>Brown University, 45 Prospect Street, Providence, RI 02912, United States</i>
rdecker@hartford.edu	<i>University of Hartford, 200 Bloomfield Ave, West Hartford, CT 06117, United States</i>
demirkaya@hartford.edu	<i>University of Hartford, Dano Hall 210, 200 Bloomfield Ave, West Hartford, CT 06117, United States</i>
junhwan.ryu@yale.edu	<i>Yale University, 206 Elm Street #205495, New Haven, CT 06520-5495, United States</i>



# The $h$ -vectors of PS ear-decomposable graphs

Nima Imani, Lee Johnson, Mckenzie Keeling-Garcia,  
Steven Klee and Casey Pinckney

(Communicated by Kenneth S. Berenhaut)

We consider a family of simple graphs known as PS ear-decomposable graphs. These graphs are one-dimensional specializations of the more general class of PS ear-decomposable simplicial complexes, which were by Chari as a means of understanding matroid simplicial complexes. We outline a shifting algorithm for PS ear-decomposable graphs that allows us to explicitly show that the  $h$ -vector of a PS ear-decomposable graph is a pure  $\mathbb{O}$ -sequence.

## 1. Introduction

This paper concerns the combinatorial structure of a certain family of simple graphs known as *PS ear-decomposable* graphs. PS ear-decomposable graphs and, more generally, PS ear-decomposable simplicial complexes, were introduced by Chari [1997] and provide a unified framework for proving a number of combinatorial results about the combinatorial structure of matroid simplicial complexes.

Stanley [1977] conjectured that the  $h$ -vector of a matroid simplicial complex is a pure  $\mathbb{O}$ -sequence. Broadly speaking, the  $h$ -vector of a graph (or more generally a simplicial complex) encodes combinatorial information about its number of vertices and edges (respectively, the number of vertices, edges, and higher-dimensional faces in a simplicial complex), and a (pure)  $\mathbb{O}$ -sequence is the degree sequence of a (pure) family of monomials that is closed under divisibility. Thus Stanley's conjecture would impose extra structure on the number of vertices and edges that a graph in this family can have (or the number of vertices, edges, and higher-dimensional faces for the family of simplicial complexes).

Chari proved that all matroid simplicial complexes are PS ear-decomposable and used this extra structure to prove a number of results on  $h$ -vectors of matroid complexes. Thus it seems natural to conjecture that the  $h$ -vector of a PS ear-decomposable simplicial complex is a pure  $\mathbb{O}$ -sequence [Chari 1997, Conjecture 3],

---

*MSC2010*: primary 05E40, 05E45; secondary 05C75.

*Keywords*: matroid,  $\mathbb{O}$ -sequence, multicomplex, ear-decomposition.

meaning that Stanley's conjecture would hold for this larger class of simplicial complexes.

In this paper, we focus our attention on the family of PS ear-decomposable graphs, which contains the family of all rank-2 matroids. The family of rank-2 matroids corresponds exactly to the family of complete multipartite graphs; but, as we will see, the family of PS ear-decomposable graphs is considerably larger. For any PS ear-decomposable graph  $\Gamma$ , we will define a canonical PS ear-decomposable graph  $\mathcal{S}(\Gamma)$  with the same number of vertices and edges as  $\Gamma$ , called a *shifted* PS ear-decomposable graph. Having defined this shifted PS ear-decomposable graph, it will be easy to find a corresponding pure multicomplex whose  $F$ -vector is the  $h$ -vector of  $\mathcal{S}(\Gamma)$ . This approach of defining a shifting algorithm as a means of preserving combinatorial data while simplifying the algebraic or geometric structure of a simplicial complex is not new, and we refer to [Kalai 2002] and the references therein for further information. It is our hope that the shifting approach presented in this paper could be generalized to higher-dimensional PS ear-decomposable simplicial complexes as an alternative approach to solving Stanley's conjecture.

## 2. Background and definitions

We will be interested in studying two families of combinatorial objects in this paper. The first is the family of PS ear-decomposable graphs, and the second is the family of pure multicomplexes.

**2.1. Graphs and PS ear-decompositions.** In this paper we only consider finite, simple graphs, which we typically denote by  $\Gamma$ . The most natural combinatorial data that can be counted for a graph  $\Gamma$  are its number of vertices and edges, which we denote by  $f_0(\Gamma)$  and  $f_1(\Gamma)$  respectively. Here the subscripts indicate that a vertex is zero-dimensional and an edge is one-dimensional when we draw a graph. We are interested in studying certain integer linear transformations of these numbers, which are called the  *$h$ -numbers* of  $\Gamma$ . The  $h$ -numbers are defined by

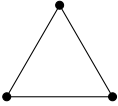
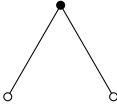
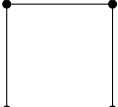

$$h_0(\Gamma) = 1, \quad h_1(\Gamma) = f_0(\Gamma) - 2, \quad h_2(\Gamma) = f_1(\Gamma) - f_0(\Gamma) + 1.$$

Notice that  $f_1(\Gamma) = h_0(\Gamma) + h_1(\Gamma) + h_2(\Gamma)$  and  $f_0(\Gamma) = h_1(\Gamma) + 2$ , so knowing the  $h$ -numbers of  $\Gamma$  is equivalent to knowing the number of vertices and edges in  $\Gamma$ . We encode the  $h$ -numbers of  $\Gamma$  in a vector called the  *$h$ -vector*, which is defined as  $h(\Gamma) = (h_0(\Gamma), h_1(\Gamma), h_2(\Gamma))$ .

Following [Chari 1997], we will study a certain family of simple graphs known as PS<sup>1</sup> ear-decomposable graphs, which are defined inductively as follows.

---

<sup>1</sup>Chari chose the name "PS ear-decomposable simplicial complexes" because products of simplices and their boundaries are fundamental to the construction.

PS cycle	$h$ -vector	PS ear	$h$ -vector contribution
	(1, 1, 1)	Type 1 	(0, 1, 1)
	(1, 2, 1)	Type 2 	(0, 0, 1)

**Table 1.** PS cycles and ears.

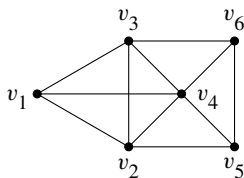
A *PS cycle* is a graph that is either a 3-cycle or a 4-cycle. A *PS ear* is a graph that is either a path of length two or a path of length one (a single edge). We call these *PS ears of Type 1* and *PS ears of Type 2* respectively. The *boundary* of a PS ear is defined as the set of vertices that are only incident to a single edge. It may seem counterintuitive to define an ear of Type 1 as a path of length two and an ear of Type 2 as a path of length one, but it will be more natural to consider ears of Type 1 first in our constructions later in the paper. Table 1 illustrates all possible PS cycles and PS ears. When illustrating PS ear-decompositions of graphs, we will adopt the practice of drawing the boundary vertices of a PS ear as unfilled circles and drawing all other vertices as filled circles.

**Definition 2.1** [Chari 1997, Section 3.3]. A graph  $\Gamma$  is *PS ear-decomposable* if it can be decomposed as a union of the form  $\Gamma = \Sigma_0 \cup \Sigma_1 \cup \dots \cup \Sigma_m$ , such that

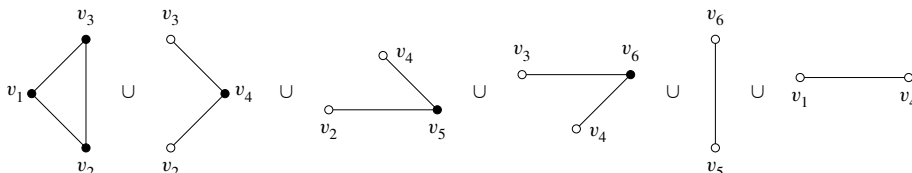
- (1)  $\Sigma_0$  is a PS cycle,
- (2)  $\Sigma_j$  is a PS ear for all  $0 < j \leq m$ , and
- (3) the intersection  $\Sigma_j \cap \bigcup_{i < j} \Sigma_i$  consists precisely of the boundary vertices of  $\Sigma_j$  for all  $0 < j \leq m$ .

One advantage to studying PS ear-decomposable graphs is that their  $h$ -vectors can also be computed inductively in terms of the ears of the decomposition. Specifically, adding an ear of Type 1 adds one vertex and two new edges to the graph, so it contributes (0, 1, 1) to the  $h$ -vector. Similarly, adding an ear of Type 2 adds one edge and zero vertices to the graph, so it contributes (0, 0, 1) to the  $h$ -vector.

**Example 2.2.** Consider the graph  $\Gamma$  on the top of page 746.



We exhibit the PS ear-decomposition of  $\Gamma$ .



Since  $\Gamma$  has 6 vertices and 11 edges, we can directly compute  $h(\Gamma) = (1, 4, 6)$ . We can also compute  $h(\Gamma)$  in terms of the given PS ear-decomposition:

$$h(\Gamma) = (1, 1, 1) + (0, 1, 1) + (0, 1, 1) + (0, 1, 1) + (0, 0, 1) + (0, 0, 1) = (1, 4, 6).$$

We note that not all graphs are PS ear-decomposable (e.g., a tree or a graph containing an induced cycle of length at least five), and some graphs may admit several combinatorially distinct PS ear-decompositions. The family of graphs that are matroid simplicial complexes is precisely the family of complete multipartite graphs, while the family of PS ear-decomposable graphs is larger, as is exhibited in Example 2.2. Furthermore, any PS ear-decomposable graph is 2-connected, and a classical theorem of Whitney [1932, Theorem 19] states that any 2-connected graph admits an *ear-decomposition*. This definition extends that of a PS ear-decomposition by allowing one to begin with a cycle of arbitrary length (not just a 3-cycle or 4-cycle) and inductively attach paths of arbitrary length (not just paths of length one or two) along their boundary vertices. Thus the family of PS ear-decomposable graphs properly contains the family of all rank-2 matroids, and is properly contained within the family of all 2-connected graphs.

**2.2. Multicomplexes.** A collection of monomials  $\mathcal{M}$  in the variables  $x_0, x_1, \dots, x_m$  is called a *multicomplex* if, whenever  $\mu \in \mathcal{M}$  and  $v$  divides  $\mu$ , then  $v \in \mathcal{M}$  as well. The name multicomplex comes from the fact that a simplicial complex is a family of sets that is closed under inclusion, so a multicomplex is a multiset analog of a simplicial complex. We refer to [Stanley 1996, Section II.2] for more information.

We say that a multicomplex  $\mathcal{M}$  has *rank*  $d$  if  $d$  is the maximal degree of any monomial in  $\mathcal{M}$ . A multicomplex  $\mathcal{M}$  is *pure of rank*  $d$  if each monomial in  $\mathcal{M}$  divides into some monomial of degree  $d$  in  $\mathcal{M}$ .

For a given multicomplex  $\mathcal{M}$  of rank  $d$ , we gather combinatorial data on  $\mathcal{M}$  in the form of the *F-vector*, written  $F(\mathcal{M}) = (F_0(\mathcal{M}), F_1(\mathcal{M}), \dots, F_d(\mathcal{M}))$ , where

degree	monomials					
2	$x_0^2$	$x_1^2$	$x_2^2$	$x_3^2$	$x_0x_1$	$x_0x_2$
1	$x_0$	$x_1$	$x_2$	$x_3$		
0	1					

**Table 2.** A pure multicomplex with  $F$ -vector  $(1, 4, 6)$ .

$F_j(\mathcal{M})$  counts the number of monomials of degree  $j$  in  $\mathcal{M}$ . An integer vector  $\mathbf{F} = (F_0, F_1, \dots, F_d)$  is a (pure)  $\mathbb{O}$ -sequence if there is a (pure) multicomplex  $\mathcal{M}$  such that  $\mathbf{F} = F(\mathcal{M})$ .

**Example 2.3.** The vector  $\mathbf{F} = (1, 3, 1)$  is an  $\mathbb{O}$ -sequence, but not a pure  $\mathbb{O}$ -sequence. The multicomplex  $\mathcal{M} = \{1, x_0, x_1, x_2, x_0x_1\}$  has  $F$ -vector  $F(\mathcal{M}) = (1, 3, 1)$ , but  $\mathbf{F}$  is not a pure  $\mathbb{O}$ -sequence since a pure multicomplex with one monomial of degree two supports at most two monomials of degree one.

**Example 2.4.** The vector  $(1, 4, 6)$  is a pure  $\mathbb{O}$ -sequence. Table 2 exhibits a pure multicomplex whose  $F$ -vector is  $(1, 4, 6)$ .

### 3. $h$ -vectors of PS ear-decomposable graphs

Stanley [1977] conjectured that the  $h$ -vector of any matroid simplicial complex is a pure  $\mathbb{O}$ -sequence. We will not define matroid simplicial complexes or their  $h$ -vectors here, but we refer to [Stanley 1996] for further details. Chari [1997] proved that any matroid simplicial complex is PS ear-decomposable, a definition that specializes to the given Definition 2.1 for graphs. Our main contribution in this paper is to show that Stanley’s conjecture continues to hold for PS ear-decomposable graphs.

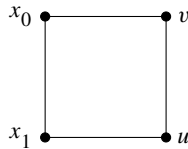
**Theorem 3.1.** *Let  $\Gamma$  be a PS ear-decomposable graph on  $n + 3$  vertices. Then there is a pure multicomplex  $\mathcal{M}$  such that  $h(\Gamma) = F(\mathcal{M})$ . Moreover, there is a canonical PS ear-decomposable graph  $\mathcal{S}(\Gamma)$  such that*

- (1)  $h(\Gamma) = h(\mathcal{S}(\Gamma))$ ,
- (2) the vertices of  $\mathcal{S}(\Gamma)$  are labeled as  $\{u, v, x_0, x_1, \dots, x_n\}$ , and
- (3) the multicomplex  $\mathcal{M}$  arises naturally from the PS ear-decomposition of  $\mathcal{S}(\Gamma)$  as a pure multicomplex on  $\{x_0, x_1, \dots, x_n\}$ .

*Proof.* We will prove Theorem 3.1 in two main steps. The first step is motivated by the observation that the  $h$ -vector of a PS ear-decomposable graph  $\Gamma$  depends only on the types of ears that are used in the PS ear-decomposition of  $\Gamma$  and is independent of the how these ears are attached. We begin by defining the graph  $\mathcal{S}(\Gamma)$ , which we call a *shifted PS ear-decomposable graph*.

Let  $\Gamma$  be a PS ear-decomposable graph on  $n + 3$  vertices with PS ear-decomposition  $\Gamma = \Sigma_0 \cup \Sigma_1 \cup \dots \cup \Sigma_m$ . For any  $0 < j < m$ , let  $\Gamma_j := \Sigma_0 \cup \Sigma_1 \cup \dots \cup \Sigma_j$ . We define a new PS ear-decomposable graph  $\mathcal{S}(\Gamma)$  satisfying conditions (1) and (2) of Theorem 3.1 by induction on the number of ears in the PS ear-decomposition of  $\Gamma$ .

If  $\Sigma_0$  is a 3-cycle, we define  $\mathcal{S}(\Gamma)_0$  to be a 3-cycle whose vertices are labeled  $u, v$ , and  $x_0$ . On the other hand, if  $\Sigma_0$  is a 4-cycle, we define  $\mathcal{S}(\Gamma)_0$  to be 4-cycle whose vertices are cyclically labeled  $u, v, x_0$ , and  $x_1$  as follows.



For  $0 < j \leq m$ , suppose we have inductively constructed a PS ear-decomposable graph  $\mathcal{S}(\Gamma)_{j-1}$  that satisfies conditions (1) and (2) of Theorem 3.1. Suppose the vertices of  $\mathcal{S}(\Gamma)_{j-1}$  are labeled as  $\{u, v, x_0, x_1, \dots, x_i\}$ . If  $\Sigma_j$  is a PS ear of Type 1, we obtain  $\mathcal{S}(\Gamma)_j$  from  $\mathcal{S}(\Gamma)_{j-1}$  by adding a new vertex labeled  $x_{i+1}$  that is adjacent to vertices  $u$  and  $v$ . Otherwise, if  $\Sigma_j$  is a PS ear of Type 2, observe that there is a missing edge in  $\mathcal{S}(\Gamma)_{j-1}$  because (i)  $\mathcal{S}(\Gamma)_{j-1}$  has the same number of vertices and edges as  $\Gamma_{j-1}$  and (ii)  $\Gamma_j$  is obtained from  $\Gamma_{j-1}$  by adding a single edge. To form  $\mathcal{S}(\Gamma)_j$ , we add the lexicographically smallest missing edge to  $\mathcal{S}(\Gamma)_{j-1}$  according to the alphabet order  $u < v < x_0 < x_1 < \dots < x_n$ . Recall that an edge  $\{a, b\}$  with  $a < b$  precedes an edge  $\{c, d\}$  with  $c < d$  lexicographically if either  $a < c$ , or  $a = c$  and  $b < d$ . By our construction it is clear that  $h(\Gamma_j) = h(\mathcal{S}(\Gamma)_j)$ .

In order to complete the proof of Theorem 3.1, we need to show that  $h(\mathcal{S}(\Gamma))$  is a pure  $\mathbb{O}$ -sequence. Again, this will follow by induction on the number of ears in the PS ear-decomposition of  $\Gamma$ . For each  $0 \leq j \leq m$ , we will construct a pure multicomplex  $\mathcal{M}_j$  such that  $F(\mathcal{M}_j) = h(\mathcal{S}(\Gamma)_j)$ .

We begin with the PS cycle  $\Sigma_0$ . If  $\Sigma_0$  is a 3-cycle, then  $h(\Sigma_0) = (1, 1, 1)$ , which is the  $F$ -vector of the pure multicomplex  $\mathcal{M}_0 = \{1, x_0, x_0^2\}$ . On the other hand, if  $\Sigma_0$  is a 4-cycle, then  $h(\Sigma_0) = (1, 2, 1)$ , which is the  $F$ -vector of the pure multicomplex  $\mathcal{M}_0 = \{1, x_0, x_1, x_0x_1\}$ .

Inductively, for  $0 < j \leq m$ , suppose we have constructed a pure multicomplex  $\mathcal{M}_{j-1}$  on variables  $\{x_0, \dots, x_i\}$  such that  $F(\mathcal{M}_{j-1}) = h(\mathcal{S}(\Gamma)_{j-1})$ . We define a pure multicomplex  $\mathcal{M}_j$  such that  $F(\mathcal{M}_j) = h(\mathcal{S}(\Gamma)_j)$  as follows:

- (1) If  $\Sigma_j$  is a PS ear of Type 1, define  $\mathcal{M}_j := \mathcal{M}_{j-1} \cup \{x_{i+1}, x_{i+1}^2\}$ . Clearly  $F(\mathcal{M}_j) = F(\mathcal{M}_{j-1}) + (0, 1, 1)$ , and hence  $h(\mathcal{S}(\Gamma)_j) = F(\mathcal{M}_j)$ . Moreover, it is clear that  $\mathcal{M}_j$  is a pure multicomplex since  $\mathcal{M}_{j-1}$  was a pure multicomplex, and we have added a new monomial of degree one and its square.

(2) If  $\Sigma_j$  is a PS ear of Type 2, define  $\mathcal{M}_j := \mathcal{M}_{j-1} \cup \mathcal{X}$ , where we define  $\mathcal{X}$  according to the following rule.

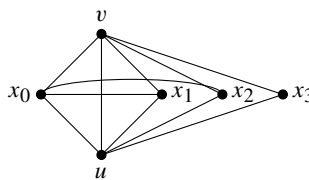
- (a) If the missing edge added to  $\mathcal{S}(\Gamma)_{j-1}$  has the form  $\{x_k, x_\ell\}$ , then  $\mathcal{X} := \{x_k x_\ell\}$ . In this case,  $\mathcal{M}_j$  is a multicomplex because the monomials of degree one that divide  $x_k x_\ell$ , which are  $x_k$  and  $x_\ell$ , belong to  $\mathcal{M}_{j-1}$  by construction; and  $\mathcal{M}_j$  is pure because we have simply added another monomial of maximal degree.
- (b) If the missing edge added to  $\mathcal{S}(\Gamma)_{j-1}$  is  $\{u, x_0\}$ , then  $\mathcal{X} := \{x_0^2\}$ ; if the missing edge is  $\{v, x_1\}$ , then  $\mathcal{X} := \{x_1^2\}$ . This only arises in the case that  $\Sigma_0$  is a 4-cycle. The monomials  $x_0^2$  and  $x_1^2$  do not belong to  $\mathcal{M}_0$  in this case, but their divisors,  $x_0$  and  $x_1$  respectively, do. Thus  $\mathcal{M}_j$  is a multicomplex, and it is pure because we have only added a monomial of maximal degree to  $\mathcal{M}_{j-1}$ .

In either case, it is again clear that  $F(\mathcal{M}_j) = F(\mathcal{M}_{j-1}) + (0, 0, 1)$  so  $h(\mathcal{S}(\Gamma)_j) = F(\mathcal{M}_j)$ .

This construction of the resulting pure multicomplex  $\mathcal{M}$  is well-defined because we do not allow multiple edges in our graphs. In the case that  $\Sigma_0$  is a 3-cycle, a monomial  $x_k^2$  is introduced when the corresponding vertex labeled  $x_k$  is introduced, and this only happens when an ear of Type 1 is attached. Otherwise, all other monomials that are introduced have the form  $x_k x_\ell$  with  $k \neq \ell$ , and correspond to an edge  $\{x_k, x_\ell\}$  being introduced to the graph. The same argument applies when  $\Sigma_0$  is a 4-cycle except that  $x_0^2$  and  $x_1^2$  are introduced to the multicomplex when the edges  $\{v, x_0\}$  and  $\{u, x_1\}$  are introduced. □

Here, we say that the graph  $\mathcal{S}(\Gamma)$  is *shifted* for the following reason. At each step in the PS ear-decomposition, an ear is attached in such a way that its boundary vertices are the lexicographically smallest pair of vertices that support the required type of ear when we order the vertices  $u < v < x_0 < \dots < x_n$ .

**Example 3.2.** Let  $\Gamma$  be the PS ear-decomposable graph presented in Example 2.2. The shifted PS ear-decomposable graph  $\mathcal{S}(\Gamma)$  is shown in Figure 2. We exhibit the PS ear-decomposition outlined in Theorem 3.1, as well as the corresponding pure multicomplex encoded by  $\mathcal{S}(\Gamma)$  in Figure 3.



**Figure 2.** The shifted graph  $\mathcal{S}(\Gamma)$ .

Ears	
Monomials	$\{1, x_0, x_0^2\} \cup \{x_1, x_1^2\} \cup \{x_2, x_2^2\} \cup \{x_3, x_3^2\} \cup \{x_0x_1\} \cup \{x_0x_2\}$

**Figure 3.** Decomposing the shifted graph  $\mathcal{G}(\Gamma)$ .

### References

[Chari 1997] M. K. Chari, “Two decompositions in topological combinatorics with applications to matroid complexes”, *Trans. Amer. Math. Soc.* **349**:10 (1997), 3925–3943. MR 98g:52023 Zbl 0889.52013

[Kalai 2002] G. Kalai, “Algebraic shifting”, pp. 121–163 in *Computational commutative algebra and combinatorics* (Osaka, 1999), edited by H. Takayuki, Adv. Stud. Pure Math. **33**, Math. Soc. Japan, Tokyo, 2002. MR 2003e:52024 Zbl 1034.57021

[Stanley 1977] R. P. Stanley, “Cohen–Macaulay complexes”, pp. 51–62 in *Higher combinatorics* (Berlin, 1976), edited by M. Aigner, NATO Adv. Study Inst. Ser., Ser. C: Math. and Phys. Sci. **31**, Reidel, Dordrecht, 1977. MR 58 #28010 Zbl 0376.55007

[Stanley 1996] R. P. Stanley, *Combinatorics and commutative algebra*, 2nd ed., Progress in Mathematics **41**, Birkhäuser, Boston, 1996. MR 98h:05001 Zbl 0838.13008

[Whitney 1932] H. Whitney, “Non-separable and planar graphs”, *Trans. Amer. Math. Soc.* **34**:2 (1932), 339–362. MR 1501641 Zbl 0004.13103

Received: 2013-06-29      Revised: 2013-10-07      Accepted: 2013-12-23

<p>imani.nima@gmail.com</p> <p>johns193@seattleu.edu</p> <p>keelingg@seattleu.edu</p> <p>klees@seattleu.edu</p> <p>pinckne1@seattleu.edu</p>	<p><i>Department of Mathematics, University of Washington, Box 354350, Seattle, WA 98195, United States</i></p> <p><i>Department of Mathematics, Seattle University, 901 12th Avenue, Seattle, WA 98122, United States</i></p> <p><i>Department of Mathematics, Seattle University, 901 12th Avenue, Seattle, WA 98122, United States</i></p> <p><i>Department of Mathematics, Seattle University, 901 12th Avenue, Seattle, WA 98122, United States</i></p> <p><i>Department of Mathematics, Seattle University, 901 12th Avenue, Seattle, WA 98122, United States</i></p>
--	---



# Zero-inflated Poisson (ZIP) distribution: parameter estimation and applications to model data from natural calamities

Sadie Beckett, Joshua Jee, Thapelo Ncube, Sophia Pompilus,  
Quintel Washington, Anshuman Singh and Nabendu Pal

(Communicated by Sat N. Gupta)

This work deals with estimation of parameters of a zero-inflated Poisson (ZIP) distribution as well as using it to model some natural calamities' data. First, we compare the maximum likelihood estimators (MLEs) and the method of moments estimators (MMEs) in terms of standardized bias (SBias) and standardized mean squared error (SMSE). We then proceed to show how datasets from some recent natural disasters can be modeled by the ZIP distribution.

## 1. Introduction

A random variable  $X$  following the usual Poisson distribution with parameter  $\lambda$ ,  $\text{Poi}(\lambda)$ , with the probability mass function

$$P(X = k) = \exp(-\lambda) \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, 3, \dots \quad (1-1)$$

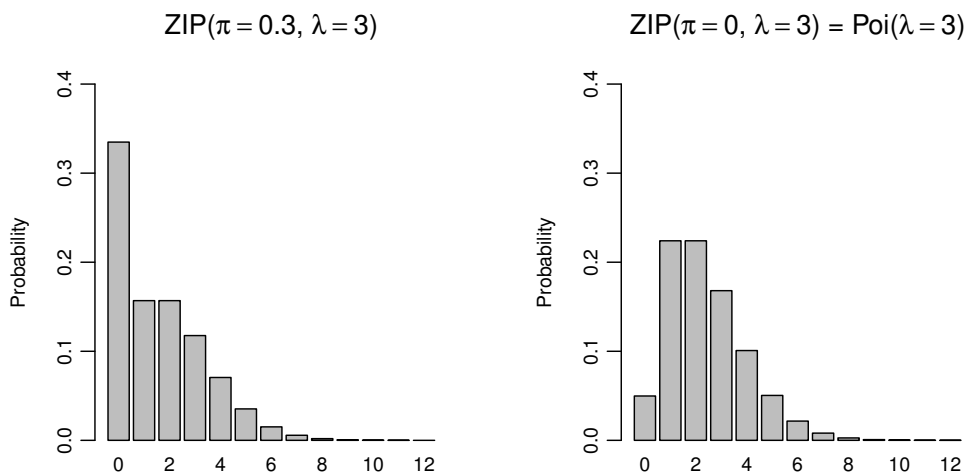
is widely used to model many naturally occurring events where  $X$  represents the "number of events per unit of time or space". Note that  $X$  takes only nonnegative integer values. However, the  $\text{Poi}(\lambda)$  distribution may not be useful (or it gives a bad fit) when  $X$  takes the value 0 with a high probability. In such a case a modified version of a regular  $\text{Poi}(\lambda)$  distribution known as the zero-inflated Poisson (ZIP) distribution becomes useful. The ZIP distribution with parameters  $\pi$  and  $\lambda$ , denoted by  $\text{ZIP}(\pi, \lambda)$ , has the following probability mass function:

$$P(X = k) = \begin{cases} \pi + (1 - \pi) \exp(-\lambda) & \text{if } k = 0 \\ (1 - \pi) \exp(-\lambda) \lambda^k / k! & \text{if } k \in \{1, 2, \dots\}, \end{cases} \quad (1-2)$$

where  $0 \leq \pi \leq 1$  and  $\lambda \geq 0$ .

*MSC2010:* primary 62F10; secondary 62F86, 62P12.

*Keywords:* method of moments estimation, maximum likelihood estimation, bias, mean squared error.



**Figure 1.** Probability distributions of ZIP and regular Poisson.

The parameter  $\pi$  gives the extra probability thrust at the value 0; when it vanishes,  $\text{ZIP}(\pi, \lambda)$  reduces to  $\text{Poi}(\lambda)$ . Figure 1 shows visually the difference between these two distributions for selected values of  $\pi$  and  $\lambda$ .

The mean and variance of  $\text{ZIP}(\pi, \lambda)$  are

$$\begin{aligned} E(X) &= \lambda(1 - \pi), \\ V(X) &= \lambda(1 - \pi)(1 + \lambda\pi). \end{aligned} \tag{1-3}$$

For example, for  $\text{ZIP}(0.3, 3)$ , these characteristics are  $E(X) = 3(1 - 0.3) = 2.1$  and  $V(X) = 3(1 - 0.3)(1 + 3(0.3)) = 3.99$ .

In the following, we provide a brief but comprehensive literature review to show how other researchers have used the ZIP distribution to model real-life data. Other important references can be found in these papers as well.

Lambert [1992] shows how a ZIP regression is better than a Poisson regression in fitting a data set with many zeros. The dataset she uses to compare these models is the number of manufacturing defects on wiring boards. Lambert concludes that ZIP regression is a straightforward model to interpret, and is convenient to use.

The decayed, missing and filled teeth (DMFT) index is used in dental epidemiology research to measure the dental health of individuals. The study [Böhning et al. 1999] used data from Brazilian school children to determine which processes were the most beneficial in preventing dental cavities. Böhning et. al state that the Poisson model often underestimates the dispersion of the data, which is why the ZIP is used instead. The ZIP model was used in this study to account for the number of children who had a DMFT of 0 (which represents good dental health). Researchers graphed the distribution of the DMFT values of the children before

and after the preventive measures were implemented in their respective schools, in order to compare the results. Besides preventive measures, intervention effects based on the ZIP model were also discussed.

Böhning [1998] asserts that the simple Poisson distribution is oftentimes inappropriate for datasets due to the numerous zeros in the data. As an example, Böhning refers to a study done with 98 HIV-positive men that provides the number of urinary tract infections experienced by these men. When the data is seen graphically, we see a huge spike at zero. We also see that there is a lack of a good fit with the Poisson model, but a good one with the ZIP model. Thus, Böhning maintains that the ZIP is a better application when there is an inflation of zeros in the count data.

Ridout, Demetrio and Hinde [1998] argue that the Poisson model does not account for high occurrences of zeros in the dataset, and therefore a better model is needed, namely the ZIP. The ZIP distribution is a slight generalization of the Poisson model, but it gives a better fit for the extra zeros.

The research described in [Davidson 2012] relates to the recurrent colorectal adenomas and the usage of the ZIP distribution. Davidson mentions that though the Poisson distribution may be used for estimated recurrences of polyp prevention trials, the ZIP is the adequate model for dealing with an inflation of zeros. This inflation was due to the fact that a large number of patients did not have recurring adenomas after being observed and treated.

The rest of the paper is organized as follows. In Section 2 we discuss the two estimation techniques and the challenges we face in using them. Section 3 covers our comprehensive simulation study to compare the two estimation techniques in terms of standardized bias (SBias) and standardized mean squared error (SMSE). In Section 4 we present some data from natural calamities where the ZIP distribution appears to provide a better fit than the usual Poisson model.

## 2. Estimation of ZIP parameters

Assume that we have independent and identically distributed (*iid*) observations  $X_1, X_2, \dots, X_n$  from  $\text{ZIP}(\pi, \lambda)$ . Our first objective is to estimate the model parameters  $\pi$  and  $\lambda$ . We are going to follow two estimation techniques, namely the Method of Moments Estimation (MME) and the Maximum Likelihood Estimation (MLE).

**2.1. The MME estimators.** Here we obtain the estimators by equating the first two sample moments with their corresponding theoretical expressions:

$$E(X) = (1 - \pi)\lambda \approx \bar{X}, \quad (2-1)$$

$$V(X) = (1 - \pi)\lambda(1 + \pi\lambda) \approx s^2, \quad (2-2)$$

$\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$  being the sample average and  $s^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$  the sample variance.

By solving (2-1) and (2-2), the MMEs are found as

$$\hat{\lambda}_{MM} = \bar{X} + \frac{s^2}{\bar{X}} - 1 \tag{2-3}$$

and

$$\hat{\pi}_{MM} = \frac{s^2}{\bar{X}} - \frac{1}{\hat{\lambda}_{MM}} = \frac{s^2 - \bar{X}}{\bar{X}^2 + (s^2 - \bar{X})}. \tag{2-4}$$

However, it must be noted that the above estimators may have the undesirable property of being negative though the parameters are nonnegative. When  $\bar{X} > s^2$ , then  $\hat{\pi}_{MM}$  can become negative, whereas the actual parameter  $\pi$  is always between 0 and 1. Therefore we are going to modify the MME by truncating  $\hat{\pi}_{MM}$  at zero and  $\hat{\lambda}_{MM}$  at  $\bar{X}$  when  $\bar{X} \geq s^2$ . The resultant estimators are called *corrected MMEs* (CMMEs) and denoted by

$$\hat{\pi}_{MM}^c = \begin{cases} 0 & \text{if } \bar{X} \geq s^2, \\ \hat{\pi}_{MM} & \text{otherwise,} \end{cases}$$

and

$$\hat{\lambda}_{MM}^c = \begin{cases} \bar{X} & \text{if } \bar{X} \geq s^2, \\ \hat{\lambda}_{MM} & \text{otherwise.} \end{cases} \tag{2-5}$$

The above CMMEs make sense because under the ZIP model  $V(X) > E(X)$  always (see (1-3)). Therefore, it is expected to have  $s^2$  to be greater than  $\bar{X}$ . Hence, a corrective measure is taken when  $\bar{X} \geq s^2$ .

**2.2. The MLE estimators.** For *iid* observations  $\tilde{X} = (X_1, \dots, X_n)$  from ZIP( $\pi, \lambda$ ), the likelihood function  $L(\pi, \lambda | \tilde{X})$  is defined as

$$L(\pi, \lambda | \tilde{X}) = \prod_{i=1}^n P(X = X_i). \tag{2-6}$$

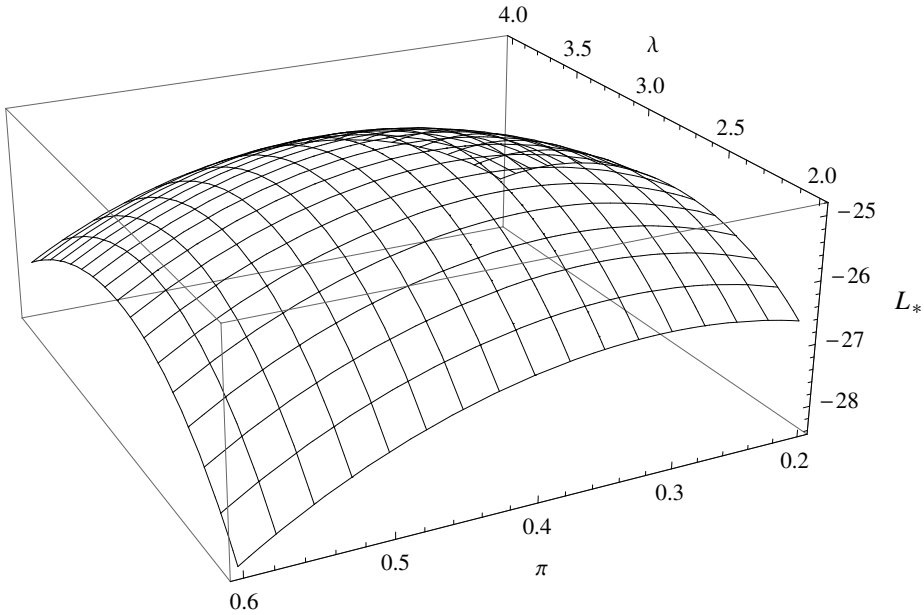
Define Y to be the number of  $X_i$ 's taking the value 0. Then

$$L(\pi, \lambda | \tilde{X}) = (\pi + (1 - \pi)e^{-\lambda})^Y \prod_{\substack{i=1 \\ X_i \neq 0}}^n (1 - \pi)e^{-\lambda} \frac{\lambda^{X_i}}{X_i!}, \tag{2-7}$$

so our log likelihood function, denoted by  $L_*$ , is

$$L_* = Y \ln(\pi + (1 - \pi)e^{-\lambda}) + (n - Y) \ln(1 - \pi) - (n - Y)\lambda + n\bar{X} \ln \lambda - \ln \prod_{i=1}^n X_i! \tag{2-8}$$

By taking the derivatives of  $L_*$  with respect to  $\pi$  and  $\lambda$ , and setting them equal to



**Figure 2.** 3D diagram of  $L_*$  plotted against  $\pi$  and  $\lambda$ .

zero, we get the following system of equations:

$$\frac{n\bar{X}}{\lambda} = \frac{Y(1 - \pi)e^{-\lambda}}{\pi + (1 - \pi)e^{-\lambda}} + n - Y, \tag{2-9}$$

$$\frac{Y(1 - \pi)(1 - e^{-\lambda})}{\pi + (1 - \pi)e^{-\lambda}} = n - Y. \tag{2-10}$$

The MLEs of  $\pi$  and  $\lambda$ , henceforth denoted by  $\hat{\pi}_{ML}$  and  $\hat{\lambda}_{ML}$  respectively, are the solutions of (2-9) and (2-10). Unlike the MMEs (or the CMMEs) we do not have explicit expressions for  $\hat{\pi}_{ML}$  and  $\hat{\lambda}_{ML}$ .

As a demonstration, we draw a random sample of size  $n = 15$  from ZIP(0.3, 3), giving us the following dataset: 0, 3, 3, 4, 0, 2, 0, 5, 0, 0, 0, 1, 3, 4, 3. The resultant log-likelihood function  $L_*$  is plotted against  $\pi$  and  $\lambda$  in Figure 2. The plot appears to have only one maximum and this has been our experience with all the replications of our simulation.

All of our computations are done using R. Widely used by statisticians, R is a free programming software for statistical computations and graphing purposes. It provides a plethora of both graphing and computational techniques, and is especially helpful for data analysis.

### 3. Comparison of two estimation techniques

In this section, we compare  $\hat{\pi}_{MM}^c$  against  $\hat{\pi}_{ML}$  and  $\hat{\lambda}_{MM}^c$  against  $\hat{\lambda}_{ML}$  in terms of *standardized bias* (SBias) and *standardized MSE* (SMSE), which are defined as

$$\text{SBias}(\hat{\theta}) = \frac{1}{\theta} \text{Bias}(\hat{\theta}) = \frac{1}{\theta} E(\hat{\theta} - \theta),$$

$$\text{SMSE}(\hat{\theta}) = \frac{1}{\theta^2} \text{MSE}(\hat{\theta}) = \frac{1}{\theta^2} E(\hat{\theta} - \theta)^2,$$

where  $\hat{\theta}$  is a generic estimator for the parameter  $\theta$ . (Note that  $\theta$  can be either  $\pi$  or  $\lambda$ , and  $\hat{\theta}$  can be the corresponding CMME or MLE.)

The usual Bias and MSE of an estimator  $\hat{\theta}$  of  $\theta$  are defined as  $\text{Bias}(\hat{\theta}) = E(\hat{\theta} - \theta)$  and  $\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$ . However, a true picture of the performance of  $\hat{\theta}$  can be judged only through SBias and/or SMSE.

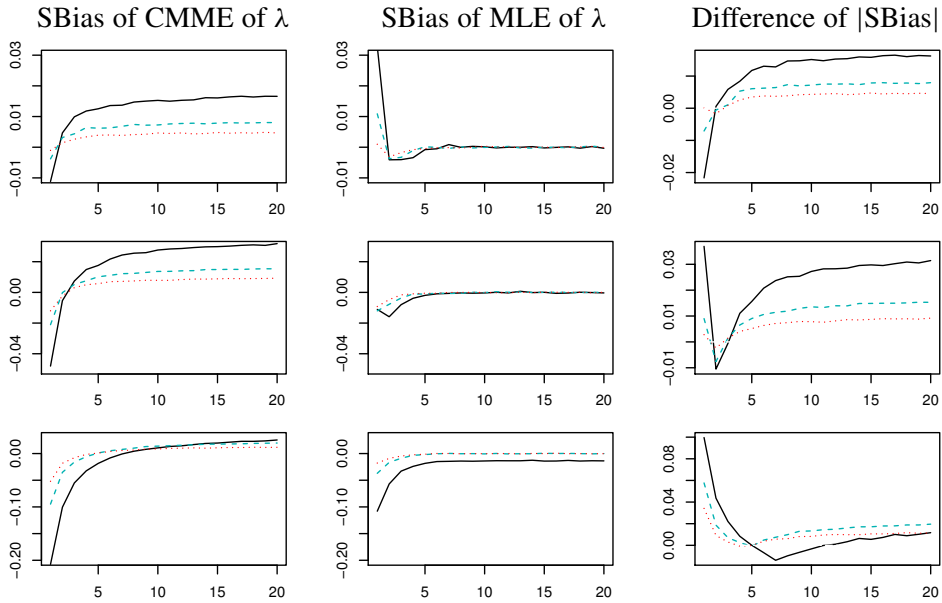
In the following, we provide the SBias and SMSE of each estimator for various values of  $n$  as well as  $(\pi, \lambda)$ . For a fixed  $n$  and  $(\pi, \lambda)$ , we generate  $X_1, \dots, X_n$  from the specified ZIP( $\pi, \lambda$ )  $10^5$  times and for each replication we compute the parameter estimates. Then an expectation is approximated by taking the average of  $N$  replicated expectants. In other words, if in the  $j$ -th replication ( $1 \leq j \leq N = 10^5$ ) we estimate a parameter  $\theta$  by  $\hat{\theta}^{(j)}$ , based on  $\tilde{X}^{(j)} = (X_1^{(j)}, X_2^{(j)}, \dots, X_n^{(j)})$ , then the SBias and SMSE are obtained by the approximations

$$\text{SBias}(\hat{\theta}) = \frac{1}{N} \sum_{j=1}^N \frac{\hat{\theta}^{(j)} - \theta}{\theta}, \quad \text{SMSE}(\hat{\theta}) = \frac{1}{N} \sum_{j=1}^N \frac{(\hat{\theta}^{(j)} - \theta)^2}{\theta^2}.$$

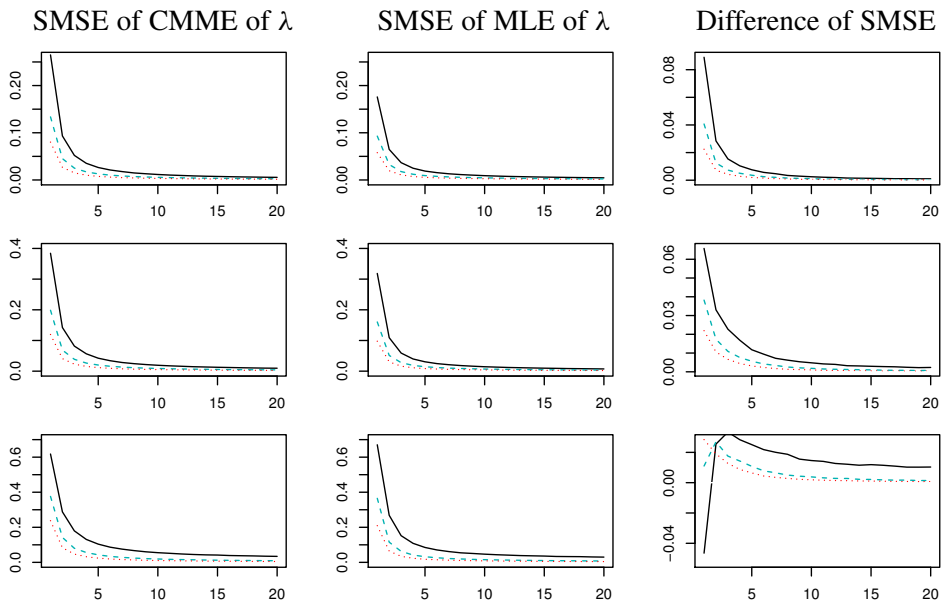
Figures 3–10 provide the plots of SBias and SMSE of estimators of  $\pi$  and  $\lambda$ . These are some of the results of our comprehensive simulation study. Every third plot across the row in each figure shows the difference between the SBias (SMSE) of CMME and SBias (SMSE) of MLE. The simulated results have been presented for small ( $n = 15$ ), moderate ( $n = 30$ ) and large ( $n = 50$ ) sample sizes. The findings of our simulation study have been summarized in the following remark.

**Remark 3.1.** For  $\lambda$  estimation, it has been observed that:

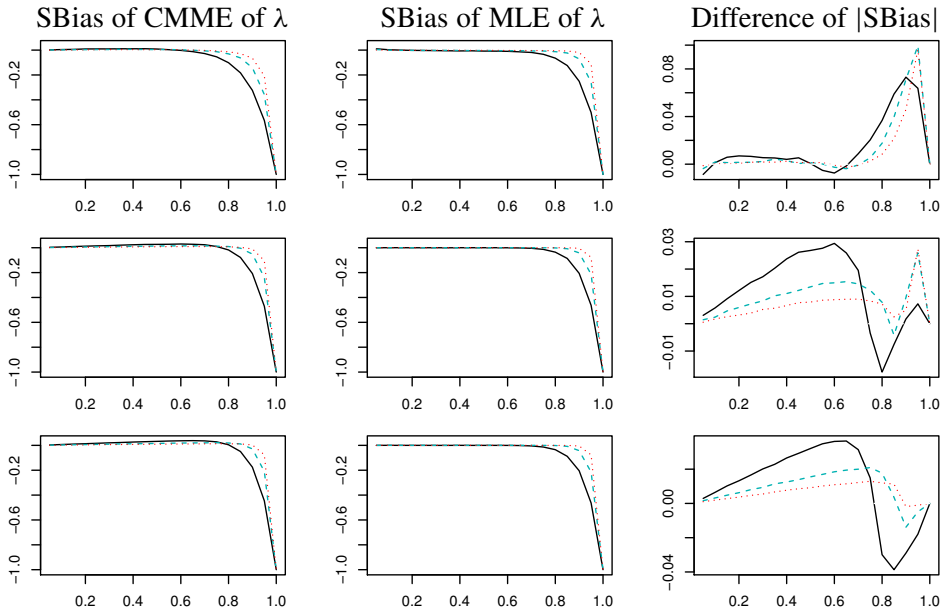
- (i)  $\hat{\lambda}_{ML}$  has mostly smaller |SBias| than that of  $\hat{\lambda}_{MM}^c$  when plotted against  $\lambda$ .
- (ii) When plotted against  $\pi$ , |SBias| of  $\hat{\lambda}_{ML}$  is smaller than that of  $\hat{\lambda}_{MM}^c$  for moderate to large sample sizes. For small  $n$ ,  $\hat{\lambda}_{ML}$  has worse |SBias| than that of  $\hat{\lambda}_{MM}^c$  over a small region of  $\pi$ .
- (iii) In terms of SMSE,  $\hat{\lambda}_{ML}$  is much superior to  $\hat{\lambda}_{MM}^c$  for all values of  $\lambda$ , except for small  $n$  when it is the other way around for small  $\lambda$ .
- (iv) The MSE of  $\lambda$  estimators, when plotted against  $\pi$ , shows superiority of  $\hat{\lambda}_{ML}$  over  $\hat{\lambda}_{MM}^c$ .



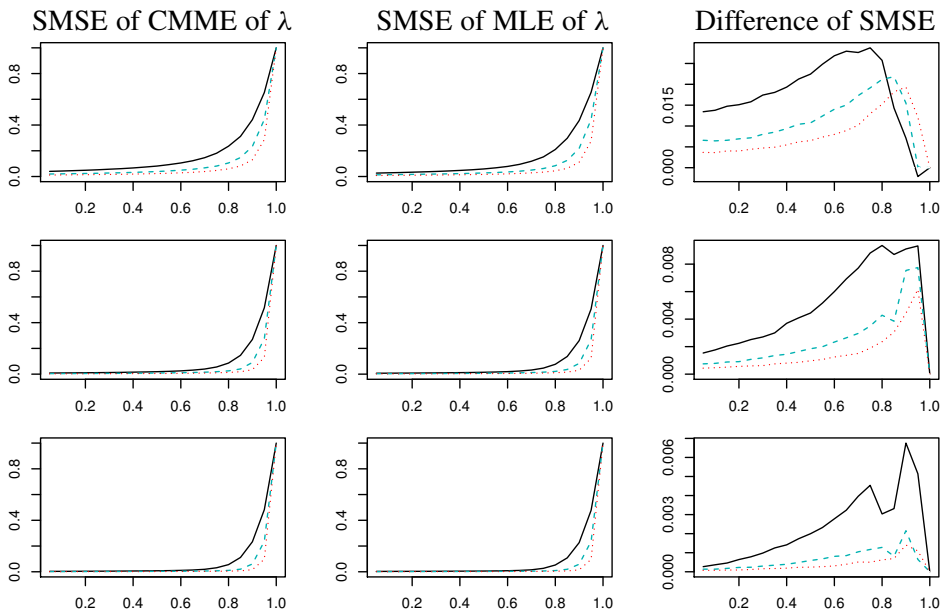
**Figure 3.** SBias study of  $\lambda$  estimators plotted against  $\lambda$ , for  $\pi = 0.25$  (top row),  $\pi = 0.5$  (middle),  $\pi = 0.75$  (bottom) and for  $n = 15$  (solid black line), 30 (dashed blue), 50 (dotted red).



**Figure 4.** SMSE study of  $\lambda$  estimators plotted against  $\lambda$ , for  $\pi = 0.25$  (top row),  $\pi = 0.5$  (middle),  $\pi = 0.75$  (bottom) and for  $n = 15$  (solid black line), 30 (dashed blue), 50 (dotted red).

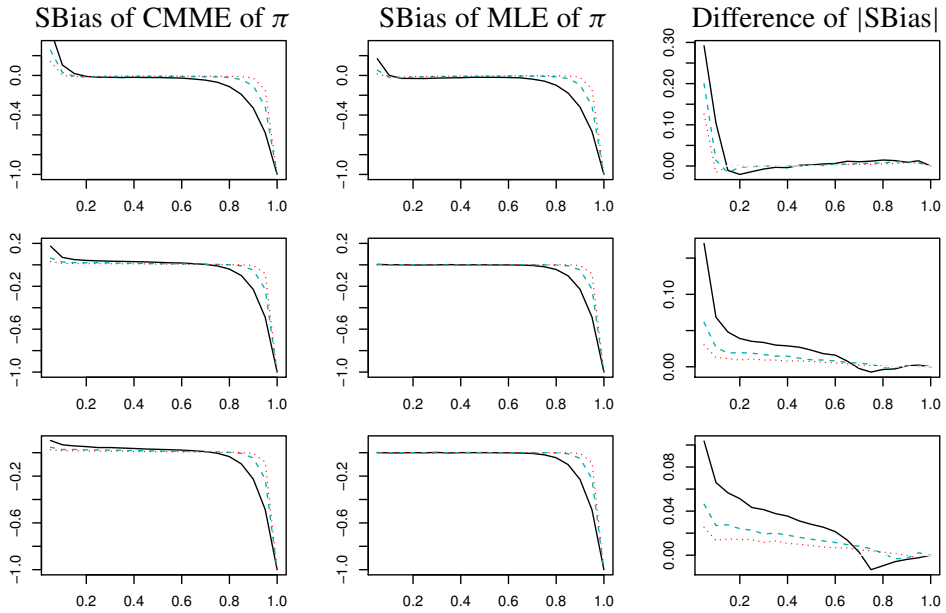


**Figure 5.** SBias study of  $\lambda$  estimators plotted against  $\pi$ , for  $\lambda = 3$  (top row),  $\lambda = 10$  (middle),  $\lambda = 50$  (bottom) and for  $n = 15$  (solid black line), 30 (dashed blue), 50 (dotted red).

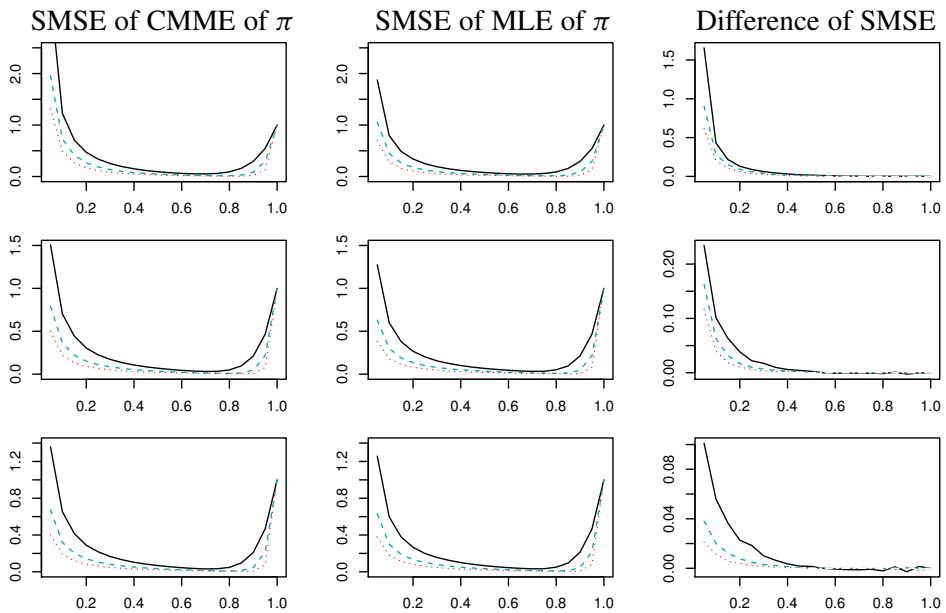


**Figure 6.** SMSE study of  $\lambda$  estimators plotted against  $\pi$ , for  $\lambda = 3$  (top row),  $\lambda = 10$  (middle),  $\lambda = 50$  (bottom) and for  $n = 15$  (solid black line), 30 (dashed blue), 50 (dotted red).

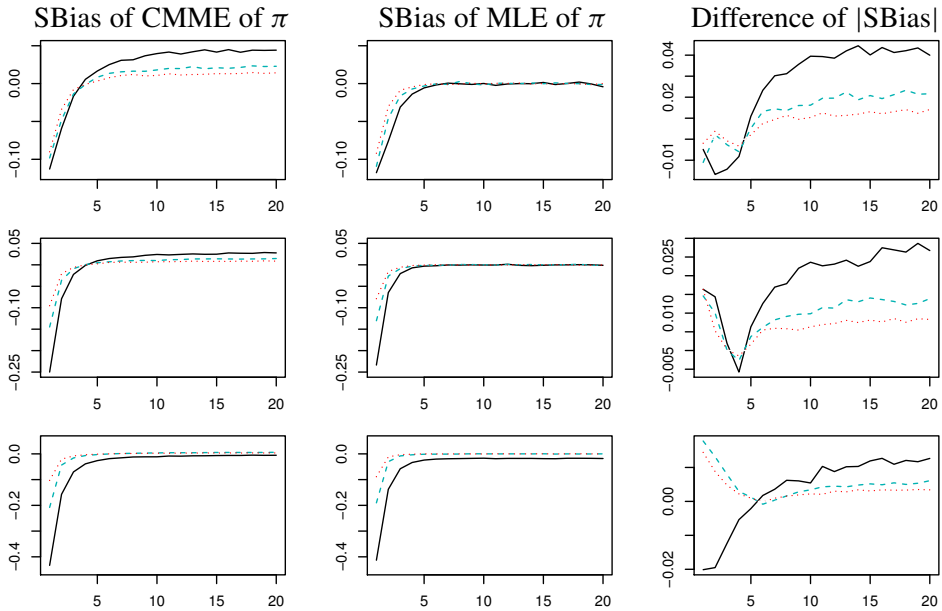




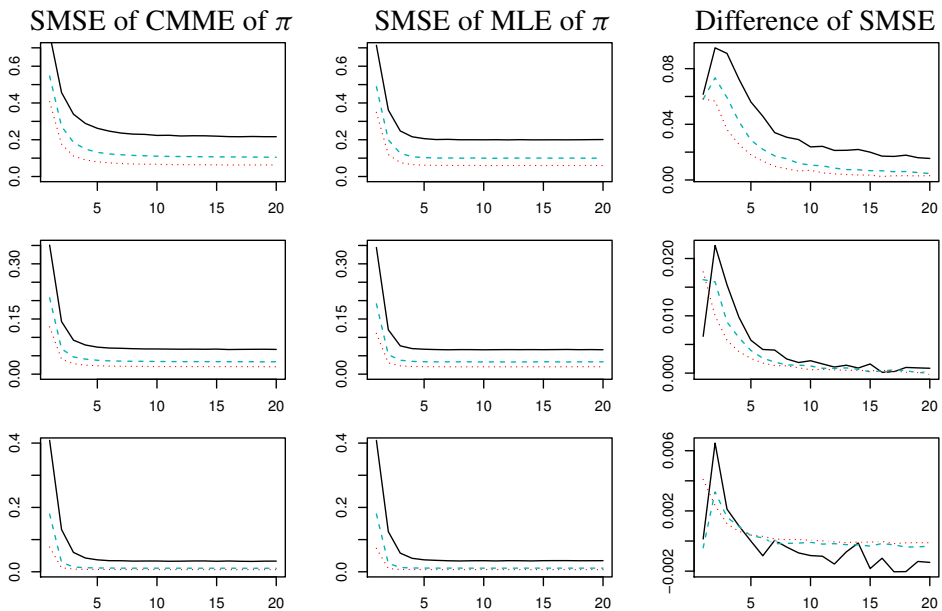
**Figure 7.** SBias study of  $\pi$  estimators plotted against  $\pi$ , for  $\lambda = 3$  (top row),  $\lambda = 10$  (middle),  $\lambda = 50$  (bottom) and for  $n = 15$  (solid black line), 30 (dashed blue), 50 (dotted red).



**Figure 8.** SMSE study of  $\pi$  estimators plotted against  $\pi$ , for  $\lambda = 3$  (top row),  $\lambda = 10$  (middle),  $\lambda = 50$  (bottom) and for  $n = 15$  (solid black line), 30 (dashed blue), 50 (dotted red).



**Figure 9.** SBias study of  $\pi$  estimators plotted against  $\lambda$ , for  $\pi = 0.25$  (top row),  $\pi = 0.5$  (middle),  $\pi = 0.75$  (bottom) and for  $n = 15$  (solid black line), 30 (dashed blue), 50 (dotted red).



**Figure 10.** SMSE study of  $\pi$  estimators plotted against  $\lambda$ , for  $\pi = 0.25$  (top row),  $\pi = 0.5$  (middle),  $\pi = 0.75$  (bottom) and for  $n = 15$  (solid black line), 30 (dashed blue), 50 (dotted red).

Similar trends hold for  $\pi$  estimators as well:

- (i) In terms of |SBias|, both  $\hat{\pi}_{MM}^c$  and  $\hat{\pi}_{ML}$  are very close for small  $\lambda$  ( $\lambda = 3$ ) when plotted against  $\pi$  but, as  $\lambda$  increases,  $\hat{\pi}_{ML}$  tends to perform better than  $\hat{\pi}_{MM}^c$  for most of  $\pi$  values.
- (ii) When |SBias| is plotted against  $\lambda$ , again  $\hat{\pi}_{ML}$  tends to perform better than  $\hat{\pi}_{MM}^c$  for most values of  $\lambda$ , especially for moderate to large sample sizes.
- (iii) In terms of SMSE, except for small  $n$ ,  $\hat{\pi}_{ML}$  performs better than  $\hat{\pi}_{MM}^c$  for most of the  $\lambda$  values.
- (iv) When SMSE is plotted against  $\pi$ ,  $\hat{\pi}_{ML}$  appears to be superior to  $\hat{\pi}_{MM}^c$  uniformly.

Based on our simulation study, the MLEs of  $\pi$  and  $\lambda$  appear to be superior estimators over their CMME counterparts, and therefore they are recommended for usage as done in Section 4 where ZIP is used to model data from natural calamities. The superiority of the MLEs has also been partially corroborated by Schwartz and Giles [2013], who observed that the MLEs exhibit very little bias even for small samples.

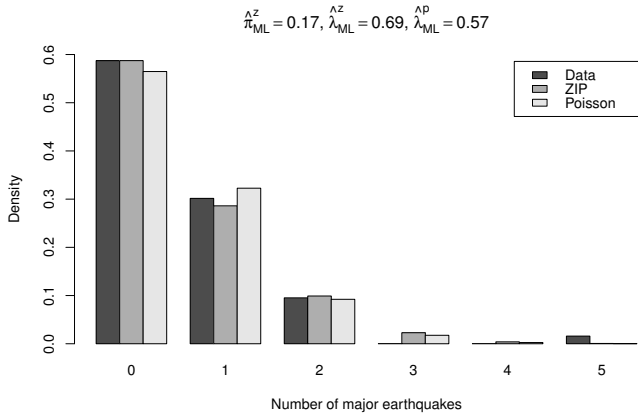
#### 4. Applications with real-life data

In this section we are going to present a few datasets from natural calamities. In each case we show the empirical probability distribution, the fitted ZIP probability distribution as well as the fitted regular Poisson probability distribution. In each of the following figures the estimated  $\lambda$  parameter under ZIP and Poisson models are denoted by  $\hat{\lambda}^z$  and  $\hat{\lambda}^p$ , respectively.

**Earthquake dataset.** Table 1 shows the number of major US earthquakes (those of magnitude at least 7.0) per year from 1950 through 2012. Figure 11 shows the plots using  $\hat{\pi}_{ML} = 0.17$  and  $\hat{\lambda}_{ML}^z = 0.69$  for ZIP and using  $\hat{\lambda}^p = 0.57$  for Poisson.

Decade	Count of yearly events									
1950–1959	0	0	1	1	1	0	0	5	2	1
1960–1969	0	0	0	0	1	2	1	0	0	0
1970–1979	0	0	1	0	0	2	0	0	0	1
1980–1989	1	0	0	0	0	0	1	1	1	0
1990–1999	0	1	2	1	1	0	1	0	0	1
2000–2009	0	0	2	2	0	1	0	1	0	0
2010–2019	0	0	0	-	-	-	-	-	-	-

**Table 1.** Number of major US earthquakes per year from 1950 through 2012 [USGS 2012].

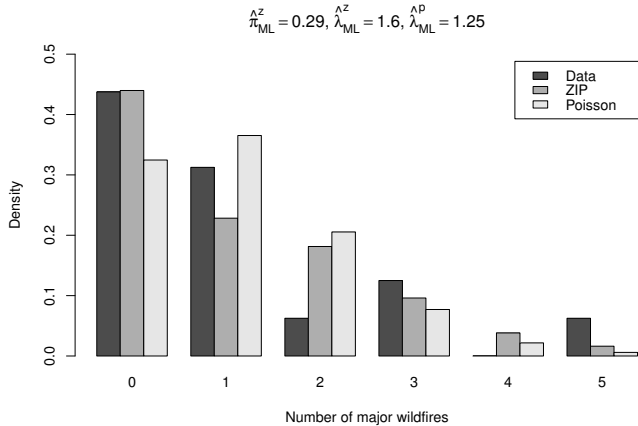


**Figure 11.** Empirical, fitted ZIP, and fitted Poisson models of the number of major earthquakes per year.

**Wildfire dataset.** Table 2 shows the number of major US wildfires (covering 400,000 acres or more) per year from 1997 through 2012. Figure 12 shows the plots using  $\hat{\pi}_{ML} = 0.29$  and  $\hat{\lambda}_{ML}^z = 1.6$  for ZIP and using  $\hat{\lambda}^p = 1.25$  for Poisson.

Decade	Count of yearly events										
1990–1999	-	-	-	-	-	-	-	-	1	0	0
2000–2009	0	0	3	0	5	1	1	1	0	3	
2010–2019	0	1	2	-	-	-	-	-	-	-	-

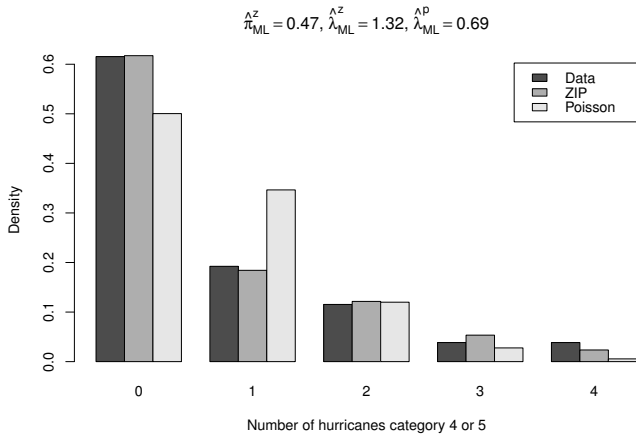
**Table 2.** Number of major US wildfires per year from 1997 through 2012 [NIFC 2012].



**Figure 12.** Empirical, fitted ZIP, and fitted Poisson models of the number of major US wildfires per year.

Decade	Count of yearly events									
1980–1989	-	-	-	-	-	-	-	0	0	1
1990–1999	0	0	1	0	0	1	0	0	2	2
2000–2009	0	0	1	1	3	4	0	0	2	0
2010–2019	0	0	0	-	-	-	-	-	-	-

**Table 3.** Number of major Atlantic hurricanes per year having landfall in the US from 1987 through 2012 [UNISYS 2012].



**Figure 13.** Empirical, fitted ZIP, and fitted Poisson models of the number of major Atlantic hurricanes per year to have landfall in the US.

**Hurricane dataset.** Table 3 shows the number of major Atlantic hurricanes (category 4 or 5) per year to have made landfall in the US from 1987 through 2012. Figure 13 shows the plots using  $\hat{\pi}_{ML} = 0.47$  and  $\hat{\lambda}_{ML}^z = 1.32$  for ZIP and using  $\hat{\lambda}^P = 0.69$  for Poisson.

**Tornado dataset.** Table 4 shows the number of tornado occurrences in Lafayette Parish, Louisiana, US per year from 1950 through 2012. Figure 14 shows the plots using  $\hat{\pi}_{ML} = 0.27$  and  $\hat{\lambda}_{ML}^z = 0.93$  for ZIP and using  $\hat{\lambda}^P = 0.63$  for Poisson.

**Lightning dataset.** Table 5 shows the number of lightning fatalities in Louisiana caused by a tree, out in the open, on golf courses, and on boats, per year from 1995 through 2012. Figure 15 shows the plots as well as estimated parameters.

**Remark 4.1.** Table 6 provides the *goodness of fit* (GOF) test results. For each dataset,  $k$  represents the number of categories (i.e., the values of  $X$ ) which is determined so that each category has at least one frequency, and the last category has been taken as  $X \geq k$ . The GOF test statistic is  $\Delta_{GOF} = \sum_{i=0}^k (O_i - E_i)^2 / E_i$ ,

Decade	Count of yearly events									
1950–1959	0	0	0	1	0	0	0	1	0	0
1960–1969	1	0	0	0	1	1	0	0	0	2
1970–1979	0	0	0	0	1	3	0	2	1	0
1980–1989	1	0	0	1	0	1	0	0	2	1
1990–1999	0	1	2	0	0	1	0	1	2	0
2000–2009	0	0	3	0	2	0	1	1	3	0
2010–2019	1	1	1	-	-	-	-	-	-	-

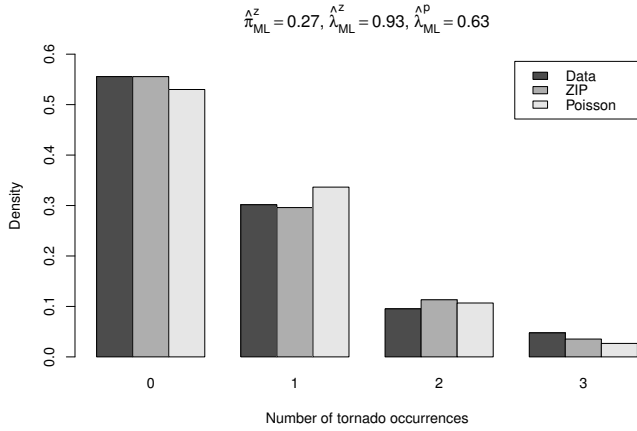
**Table 4.** Number of tornado occurrences in Lafayette Parish, Louisiana per year from 1950 through 2012 [NOAA 2012b].

Decade	Fatalities by a tree									
1990–1999	-	-	-	-	-	0	1	0	0	0
2000–2009	0	0	1	0	0	0	0	0	0	0
2010–2019	0	0	2	-	-	-	-	-	-	-
	Fatalities in the open									
1990–1999	-	-	-	-	-	1	0	0	2	1
2000–2009	0	1	1	0	0	1	0	0	0	0
2010–2019	1	0	0	-	-	-	-	-	-	-
	Fatalities on golf courses									
1990–1999	-	-	-	-	-	0	0	0	0	2
2000–2009	0	0	0	0	0	0	0	0	0	0
2010–2019	0	1	0	-	-	-	-	-	-	-
	Fatalities on boats									
1990–1999	-	-	-	-	-	0	0	0	2	1
2000–2009	2	0	0	0	1	0	0	0	0	1
2010–2019	0	0	0	-	-	-	-	-	-	-

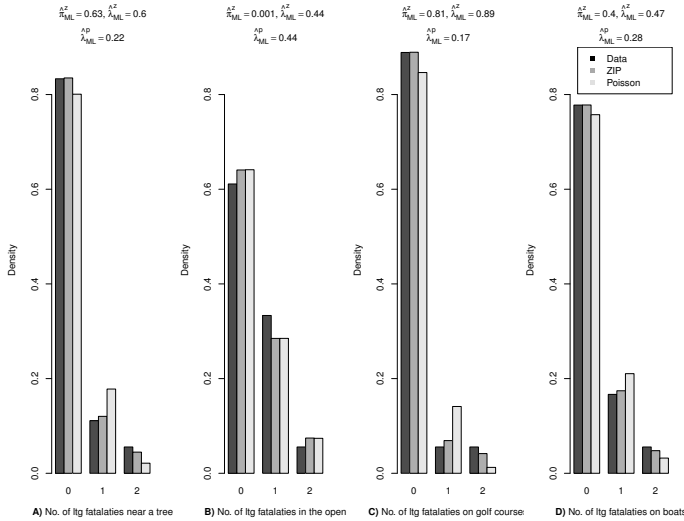
**Table 5.** Number of lightning fatalities by category per year 1995 through 2012 [NOAA 2012a].

where  $O_i$  is the observed frequency of the event  $X = i$  for  $i \leq k - 1$ , and the event  $X \geq k$  when  $i = k$ , and  $E_i$  is the expected frequency obtained by multiplying the corresponding fitted probability by the sample size  $n$ . Note that the  $p$ -values are all more than 75%, and mostly 90% or higher. This clearly shows that the ZIP model gives a very good fit to model the severe natural calamities which occur rarely.

**Remark 4.2.** In some of these plots it is seen that the fitted Poisson model comes very close to the fitted ZIP model, namely when the estimated  $\pi$  is very close to



**Figure 14.** Empirical, fitted ZIP, and fitted Poisson models of the number of tornado occurrences per year in Lafayette, Louisiana.



**Figure 15.** Empirical, fitted ZIP, and fitted Poisson models of the number of lightning fatalities in Louisiana for specified situations.

zero. It is reasonable to expect that  $\hat{\pi}_{ML}$  being close to 0 implies that  $\pi = 0$ , i.e., that the ZIP model reduces to the regular Poisson model. Currently, hypothesis testing on the ZIP parameters is under consideration, and will be reported in near future.

**Concluding remark**

Using the fitted ZIP model, one can estimate that in any year, the probability of having at least one major earthquake in the US is 0.4136 or approximately 41%.

	Observed $\Delta_{\text{GOF}}$ value $\delta$	$p$ -value = $P(\chi_{k-2}^2 > \delta)$	Conclusion
Earthquake data	$\delta = .3575, k = 4$	.9489	good fit
Wildfire data	$\delta = 1.892, k = 5$	.7556	good fit
Hurricane data	$\delta = .3680, k = 5$	.9850	good fit
Tornado data	$\delta = .3637, k = 5$	.9853	good fit
Lightning data			
outside	$\delta = .2628, k = 3$	.8769	good fit
near tree	$\delta = .0606, k = 3$	.9702	good fit
on a golf course	$\delta = .1218, k = 3$	.9409	good fit
on a boat	$\delta = .4158, k = 3$	.8123	good fit

**Table 6.** Goodness of fit results.

Similarly, the probability of Lafayette Parish getting hit by a tornado in any year is 0.4420 or approximately 44%. Hopefully these probabilities may find applications in the insurance industry, and this study will stimulate further research in this direction.

### Acknowledgements

This work is a part of the undergraduate summer research and professional development experience, organized by the University of Louisiana at Lafayette, called “Smooth Transition for Advancement to Graduate Education” (STAGE) for Underrepresented Minorities (URM) in Mathematical Sciences. STAGE is a Pilot Project supported by the National Science Foundation under the Grant DMS-1043223. Five URM Students — S. Beckett (University of Louisiana at Lafayette, LA), J. Jee (Louisiana College, Pineville, LA), T. Ncube (University of West Florida, Pensacola, FL), S. Pompilus (Georgetown University, Washington, D.C.), and Q. Washington (Southern University, New Orleans, LA) — constituted the statistics group of the STAGE 2013 program. With help from A. Singh (graduate assistant) on numerical computations and statistical simulations, the statistics group worked under the supervision of Professor N. Pal, Principal Investigator of STAGE.

### References

- [Böhning 1998] D. Böhning, “Zero-inflated Poisson models and C.A.MAN: a tutorial collection of evidence”, *Biometric Model* **40**:7 (1998), 833–843. Zbl 0914.62091
- [Böhning et al. 1999] D. Böhning, E. Dietz, P. Schlattmann, L. Mendonça, and U. Kirchner, “The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology”, *J. R. Statist. Soc. A* **162**:2 (1999), 195–209.
- [Davidson 2012] C. L. Davidson, *A model selection paradigm for modeling recurrent adenoma data in polyp prevention trials*, Master’s thesis, University of Arizona, 2012, <http://arizona.openrepository.com/arizona/handle/10150/228465>.



- [Lambert 1992] D. Lambert, “Zero-inflated Poisson regression with an application to defects in manufacturing”, *Technometrics* **34**:1 (1992), 1–14. Zbl 0850.62756
- [NIFC 2012] NIFC, “1997–2012 large fires (100,000+ acres)”, 2012, [http://www.nifc.gov/fireInfo/fireInfo\\_stats\\_lgFires.html](http://www.nifc.gov/fireInfo/fireInfo_stats_lgFires.html). Table.
- [NOAA 2012a] NOAA, “Natural hazard statistics: lightning”, 2012, <http://www.nws.noaa.gov/om/hazstats.shtml>.
- [NOAA 2012b] NOAA, “Storm prediction center warning coordination meteorologist page”, statistical report, 2012, <http://www.spc.noaa.gov/wcm/#data>.
- [Ridout et al. 1998] M. Ridout, C. G. B. Demetrio, and J. Hinde, “Models for count data with many zeros”, pp. 179–192 in *Proceedings of the XIXth International Biometric Conference* (Cape Town, 1998), International Biometric Society, Washington, DC, 1998.
- [Schwartz and Giles 2013] J. Schwartz and D. E. Giles, “Biased-reduced maximum likelihood estimation for the zero-inflated Poisson distribution”, preprint, 2013, <http://econpapers.repec.org/RePEc:vic:vicwp:1102>. To appear in *Commun. Stat. Theory Methods*.
- [UNISYS 2012] UNISYS, “Atlantic tropical storm tracking by year”, statistical report, 2012, <http://weather.unisys.com/hurricane/atlantic/index.php>.
- [USGS 2012] USGS, “Historic earthquakes in the United States and its territories”, statistical report, 2012, <http://earthquake.usgs.gov/earthquakes/states/historical.php>.

Received: 2013-07-24

Revised: 2013-12-06

Accepted: 2014-01-03

smbek17@aol.com	<i>University of Louisiana at Lafayette, 104 E University Avenue, Lafayette, LA 70504, United States</i>
joshua92891@yahoo.com	<i>Louisiana College, Pineville, LA 71360, United States</i>
tmn11@students.uwf.edu	<i>University of West Florida, 11000 University Parkway, Pensacola, FL 32514, United States</i>
sp485@georgetown.edu	<i>Georgetown University, 3700 O St NW, Washington, DC 20057, United States</i>
washingtell@gmail.com	<i>Southern University New Orleans, 6400 Press Drive, New Orleans, LA 70126, United States</i>
anshumandotsingh@gmail.com	<i>University of Louisiana at Lafayette, 104 E University Avenue, Lafayette, LA 70504, United States</i>
npx3695@louisiana.edu	<i>Department of Mathematics, University of Louisiana at Lafayette, PO Box 41010, Lafayette, LA 70504, United States</i>



# On commutators of matrices over unital rings

Michael Kaufman and Lillian Pasley

(Communicated by Chi-Kwong Li)

Let  $R$  be a unital ring and let  $X \in M_n(R)$  be any upper triangular matrix of trace zero. Then there exist matrices  $A$  and  $B$  in  $M_n(R)$  such that  $X = [A, B]$ .

## 1. Introduction

Shoda [1936] proved that every matrix with trace zero over the complex numbers could be expressed as a commutator  $AB - BA$ . Albert and Muckenhoupt [1957] extended this result to matrices over any field. For matrices over commutative rings it is known that matrices of trace zero in general cannot be presented as commutators [Lissner 1961; Rosset and Rosset 2000]. Recently, Khurana and Lam [2012] showed every matrix with trace zero over any field can be expressed as a generalized commutator  $ABC - CBA$ . But the same result does not hold for matrices over commutative rings. Our work is motivated by the following question posed by Khurana and Lam: if  $n \geq 3$ , is every upper triangular matrix a generalized commutator over any ring  $S$  [Khurana and Lam 2012, Question 8.17]. In the case when  $n = 2$  this question has a negative answer as has been shown in [Khurana and Lam 2012, Theorem 8.11]. Using ideas due to Khurana and Lam we will give a simple proof of this case. We will also show that every  $n \times n$  upper triangular matrix of trace zero over any unital ring can be presented as a commutator.

## 2. Results

In this section, the trace of an  $n \times n$  matrix  $M = (x_{i,j})$  is denoted  $\text{tr}(M) = \sum_{k=1}^n x_{k,k}$ . Let  $R$  be any ring and  $S$  any commutative ring. We need some auxiliary results.

**Proposition 1** [Khurana and Lam 2012, Proposition 6.6]. *Let*

$$X = [A, B, C] = ABC - CBA,$$

where  $X, A, B, C \in M_n(S)$ . Then  $\text{tr}(BX) = 0$ .

*MSC2010:* primary 15A54; secondary 16S50.

*Keywords:* trace, matrix algebra, unital ring.

Supported in part by National Science Foundation, grant no. 1156798.

**Proposition 2** [Khurana and Lam 2012, Proposition 8.3]. *Let  $D \in R$  such that  $DC = CD \in Z(R)$  (the center of  $R$ ). If  $X = [A, B, C] \in R$ , then*

$$DX = [D, ABC] + [A, BCD] \quad \text{and} \quad XD = [D, CBA] + [A, BCD].$$

If  $X, D \in M_n(S)$ , then  $\text{tr}(XD) = \text{tr}(DX) = 0$ .

Khurana and Lam showed for  $n \geq 2$  there exist  $n \times n$  matrices that can not be expressed as generalized commutators. Now we use the preceding propositions to provide a different proof for the  $n = 2$  case.

**Theorem 3** [Khurana and Lam 2012, Theorem 8.11]. *There exists a  $2 \times 2$  upper triangular matrix that can not be expressed as a generalized commutator (i.e.,  $X \neq ABC - CBA$ ).*

*Proof.* Let  $A = (a_{ij}), B = (b_{ij}), C = (c_{ij}), A, B, C \in M_2(S)$ , where  $S = \mathbb{C}[x, y, z]$  and  $x, y$ , and  $z$  are indeterminates. Now suppose  $X \in M_2(S)$  is the upper triangular matrix  $\begin{pmatrix} x & y \\ 0 & z \end{pmatrix}$  such that  $X = ABC - CBA$ .

We begin by observing that

$$BX = \begin{pmatrix} b_{11}x & b_{11}y + b_{12}z \\ b_{21}x & b_{21}y + b_{22}z \end{pmatrix}.$$

By Proposition 1,  $\text{tr}(BX) = b_{11}x + b_{21}y + b_{22}z = 0$ . This implies that polynomials  $b_{11}, b_{21}$ , and  $b_{22}$  cannot contain constant terms.

We consider the characteristic equation of  $A$ . From  $A^2 + \lambda A + \mu I = 0$  where

$$\lambda = -\text{tr}(A) = -a_{11} - a_{22} \quad \text{and} \quad \mu = \det(A) = a_{11}a_{22} - a_{12}a_{21},$$

we see that  $A(A + \lambda I) = -\mu I$ , and so  $A(A + \lambda I) \in Z(S)$ . Now we examine

$$(A + \lambda I)X = \begin{pmatrix} -a_{22}x & -a_{22}y + a_{12}z \\ a_{21}x & a_{21}y - a_{11}z \end{pmatrix}.$$

By Proposition 2,  $\text{tr}((A + \lambda I)X) = -a_{22}x + a_{21}y - a_{11}z = 0$ . This implies that polynomials  $a_{11}, a_{21}$ , and  $a_{22}$  cannot contain constant terms. Similarly, polynomials  $c_{11}, c_{21}$ , and  $c_{22}$  cannot contain constant terms. From  $X = ABC - CBA$  we obtain

$$x = a_{12}(b_{21}c_{11} + b_{22}c_{21}) + b_{12}(a_{11}c_{21} - a_{21}c_{11}) + c_{12}(-a_{11}b_{21} - a_{21}b_{22}). \quad (1)$$

Polynomials  $a_{11}, a_{21}, a_{22}, b_{11}, b_{21}, b_{22}, c_{11}, c_{21}$ , and  $c_{22}$  contain no constant terms, so the right-hand side of (1) cannot contain a linear term. Since the left-hand side of (1) is a polynomial of degree 1, namely  $x$ , we arrive at a contradiction.  $\square$

Since there exist upper triangular matrices in  $M_n(S)$  that cannot be expressed as generalized commutators, we consider what can be said about upper triangular matrices with respect to commutators.

**Theorem 4.** *Let  $R$  be a unital ring and let  $X \in M_n(R)$  be any upper triangular matrix of trace zero. Then there exist matrices  $A$  and  $B$  in  $M_n(R)$  such that  $X = [A, B]$ .*

This theorem is not true without the assumption that  $R$  is a unital ring. Let  $R$  be the ring of polynomials over  $\mathbb{C}$  with zero constant terms in variable  $x$ . Then

$$X = \begin{pmatrix} x & 0 & \cdots & 0 & 0 \\ 0 & x & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & x & 0 \\ 0 & 0 & \cdots & 0 & -(n-1)x \end{pmatrix}$$

is of trace zero. However, the entries of a nonzero commutator  $[A, B]$  in  $M_n(R)$  do not contain any linear terms.

*Proof of Theorem 4.* Let  $X \in M_n(R)$  be an upper triangular matrix of the form

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1,n} \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & x_{n-1,n-1} & x_{n-1,n} \\ 0 & \cdots & 0 & -\sum_{k=1}^{n-1} x_{k,k} \end{pmatrix}.$$

Let

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

We define the matrix  $B$  as follows: for  $1 \leq i-1 \leq j \leq n$ , let

$$b_{ij} = \sum_{k=1}^{i-1} x_{k,j-i+k+1}.$$

All other terms of  $B$  are zero. Our goal is to show that  $X = [A, B]$ . Let  $[A, B] = (t_{i,j})$ . We want to prove  $t_{i,j} = x_{i,j}$  for  $i \geq j$ ,

$$t_{n,n} = -\sum_{k=1}^{n-1} x_{k,k},$$

and  $t_{i,j} = 0$  for  $i < j$ . We will split the proof into four cases.

Case 1. If  $i > j$ , then  $t_{i,j} = b_{i+1,j} - b_{i,j-1} = 0$ .

Case 2. If  $i = j = 1$ , then  $t_{i,j} = b_{21} = x_{11}$ .

Case 3. If  $i = j = n$ , then

$$t_{i,j} = 0 - b_{n,n-1} = 0 - \sum_{k=1}^{n-1} x_{k,k} = - \sum_{k=1}^{n-1} x_{k,k}.$$

Case 4. If  $i < j$  or  $i = j \in \{2, 3, \dots, n-1\}$ , then

$$t_{i,j} = b_{i+1,j} - b_{i,j-1} = \sum_{k=1}^i x_{k,j-i+k} - \sum_{k=1}^{i-1} x_{k,j-i+k} = x_{i,j}.$$

This completes the proof.  $\square$

This result may be used to give a proof of the well-known theorem due to Shoda [1936].

**Corollary 5.** *Let  $\mathbb{C}$  be the field of complex numbers and  $M_n(\mathbb{C})$  be the ring of  $n \times n$  matrices. Then every matrix of trace zero is a commutator.*

*Proof.* Let  $P$  be any matrix of trace zero and  $Q$  be Jordan normal form for  $P$ . So we have  $P = C^{-1}QC$  for some invertible  $C$ . Since  $P$  is upper triangular and of trace zero by Theorem 4 there exist  $A, B \in M_n(\mathbb{C})$  such that  $Q = [A, B]$ . Therefore,  $P = C^{-1}QC = C^{-1}[A, B]C = [C^{-1}AC, C^{-1}BC]$ .  $\square$

### Acknowledgments

We are grateful to our mentor Dr. Mikhail Chebotar for his guidance in writing this paper. We would also like to thank the referee for the careful reading of this paper.

### References

- [Albert and Muckenhoupt 1957] A. A. Albert and B. Muckenhoupt, “On matrices of trace zeros”, *Michigan Math. J.* **4** (1957), 1–3. MR 18,786b Zbl 0077.24304
- [Khurana and Lam 2012] D. Khurana and T. Y. Lam, “Generalized commutators in matrix rings”, *Linear Multilinear Algebra* **60**:7 (2012), 797–827. MR 2929647 Zbl 1255.15015
- [Lissner 1961] D. Lissner, “Matrices over polynomial rings”, *Trans. Amer. Math. Soc.* **98** (1961), 285–305. MR 23 #A171 Zbl 0111.01703
- [Rosset and Rosset 2000] M. Rosset and S. Rosset, “Elements of trace zero that are not commutators”, *Comm. Algebra* **28**:6 (2000), 3059–3072. MR 2001c:16055 Zbl 0954.16021
- [Shoda 1936] K. Shoda, “Einige Sätze über Matrizen”, *Jap. J. Math.* **13** (1936), 361–365. Zbl 0017.05101

Received: 2013-08-09

Revised: 2014-03-04

Accepted: 2014-03-08

mkaufma5@kent.edu

Kent State University, Kent, OH 44240, United States

lfpasley@ncsu.edu

North Carolina State University, Raleigh, NC 27695,  
United States

# The nonexistence of cubic Legendre multiplier sequences

Tamás Forgács, James Haley, Rebecca Menke and Carlee Simon

(Communicated by Michael Dorff)

Our main result is the proof of the recently conjectured nonexistence of cubic Legendre multiplier sequences. We also give an alternative proof of the nonexistence of linear Legendre multiplier sequences using a method that will allow for a more methodical treatment of sequences interpolated by higher degree polynomials.

## 1. Introduction

Given a simple set of polynomials  $Q = \{q_k(x)\}_{k=0}^{\infty}$  and a sequence of numbers  $\{\gamma_k\}_{k=0}^{\infty}$ , one can define the operator associated with  $\{\gamma_k\}_{k=0}^{\infty}$  as  $T[q_k(x)] = \gamma_k q_k(x)$  for  $k = 0, 1, 2, \dots$ , and extend its action to  $\mathbb{R}[x]$  linearly. Our work in this paper concerns such operators when  $Q$  consists of the Legendre polynomials.

**Definition 1.** The Legendre polynomials  $\mathfrak{L}e_k(x)$  are defined by the generating relation

$$\frac{1}{\sqrt{1-2xt+t^2}} = \sum_{k=0}^{\infty} \mathfrak{L}e_k(x)t^k,$$

where the square root denotes the branch which goes to 1 as  $t \rightarrow 0$ .

**Definition 2.** A sequence of real numbers  $\{\gamma_k\}_{k=0}^{\infty}$  is a Legendre multiplier sequence if  $\sum_{k=0}^n a_k \gamma_k \mathfrak{L}e_k(x)$  has only real zeros whenever  $\sum_{k=0}^n a_k \mathfrak{L}e_k(x)$  has only real zeros. We define  $Q$ -multiplier sequences for any basis  $Q$  of  $\mathbb{R}[x]$  analogously. If  $Q$  is the standard basis, the associated multiplier sequences are called classical multiplier sequences (of the first kind).

*MSC2010:* 26C10, 30C15.

*Keywords:* Legendre multiplier sequences, reality preserving linear operators, symbol of a linear operator, coefficients of Legendre-diagonal differential operators.

Research partially supported by NSF grant DMS-1156273. Some of the work was completed while Forgács was on sabbatical leave at the University of Hawai'i at Manoa, whose support he gratefully acknowledges.

Every sequence of the form  $0, 0, 0, \dots, a, b, \dots, 0, 0, 0, \dots$ , where  $a, b \in \mathbb{R}$ , is a Legendre multiplier sequence. The literature calls such sequences *trivial*. In addition to these, there is an abundance of *nontrivial* Legendre multiplier sequences (see [Blakeman et al. 2012] for examples). Thus, the problem of characterizing polynomials which interpolate Legendre multiplier sequences is a meaningful one, and it fits well into the landscape of current research in the theory of multiplier sequences (see, for example, [Blakeman et al. 2012; Brändén and Ottergren 2014; Forgács and Piotrowski 2013b; Yoshida 2013]). The present paper contributes to this line of inquiry by settling a conjecture on the nonexistence of cubic Legendre multiplier sequences [Blakeman et al. 2012, Open problem (1)]. In addition, we give a new proof of the nonexistence of linear Legendre multiplier sequences, which is more methodical than the educated hunt for test polynomials whose zeros fail to remain real after having been acted on by a linear sequence.

The rest of the paper is organized as follows. In Section 2 we present a number of known results which are relevant to our investigations. Section 3 exhibits a new proof of the nonexistence of linear Legendre multiplier sequences [Blakeman et al. 2012, Proposition 2] using a theorem of Borcea and Brändén. Our method exploits the fact that one does not need to have full knowledge of all coefficient polynomials  $T_k(x)$  of a linear operator  $T = \sum_{k=0}^{\infty} T_k(x)D^k$  in order to decide whether or not  $T$  is reality preserving. Section 4 contains the main result, Theorem 17, which establishes the nonexistence of cubic Legendre multiplier sequences. We conclude with a section on open problems.

## 2. Background

Central to the theory of (classical) multiplier sequences is the Laguerre–Pólya class of real entire functions, which we denote by  $\mathcal{L}\text{-}\mathcal{P}$ . We recall the definition here, along with a recent theorem characterizing this class as precisely those real entire functions which satisfy the generalized Laguerre inequalities.

**Definition 3.** A real entire function  $\varphi(x) = \sum_{k=0}^{\infty} (\gamma_k/k!)x^k$  is said to belong to the Laguerre–Pólya class, written  $\varphi \in \mathcal{L}\text{-}\mathcal{P}$ , if it can be written in the form

$$\varphi(x) = cx^m e^{-ax^2+bx} \prod_{k=1}^{\omega} \left(1 + \frac{x}{x_k}\right) e^{-x/x_k},$$

where  $b, c \in \mathbb{R}$ ,  $x_k \in \mathbb{R} \setminus \{0\}$ ,  $m$  is a nonnegative integer,  $a \geq 0$ ,  $0 \leq \omega \leq \infty$  and  $\sum_{k=1}^{\omega} 1/x_k^2 < \infty$ . If  $\gamma_k \geq 0$  for all  $k = 0, 1, 2, \dots$ , we say that  $\varphi \in \mathcal{L}\text{-}\mathcal{P}^+$ .

Csordas and Vishnyakova recently completed the following characterization of the class  $\mathcal{L}\text{-}\mathcal{P}$ .



**Theorem 4** [Csordas and Varga 1990, Theorem 2.9; Csordas and Vishnyakova 2013, Theorem 2.3]. *Let  $\varphi(x)$  denote a real entire function, with  $\varphi(x) \not\equiv 0$ . Then  $\varphi \in \mathcal{L}\text{-}\mathcal{P}$  if and only if for all  $n \in \mathbb{N}_0$  and for all  $x \in \mathbb{R}$ ,*

$$L_n(x, \varphi) := \sum_{j=0}^{2n} \frac{(-1)^{j+n}}{(2n)!} \binom{2n}{j} \varphi^{(j)}(x) \varphi^{(2n-j)}(x) \geq 0.$$

We shall make use of this theorem in Section 3 when we reprove the nonexistence of linear Legendre multiplier sequences. Since  $\mathcal{L}\text{-}\mathcal{P}$  is exactly the class of real entire functions which are locally uniform limits on  $\mathbb{C}$  of real polynomials with only real zeros (see [Levin 1956, Chapter VIII] or [Obreschkoff 1963, Satz 3.2]), it is closed under differentiation. Thus if  $\varphi \in \mathcal{L}\text{-}\mathcal{P}$ , then

$$L_1(x, \varphi^{(k)}(x)) \geq 0 \quad \text{for all } k \in \mathbb{N}_0.$$

Pólya and Schur [1914] completely characterized classical multiplier sequences. Their seminal theorem maintains relevance in the setting of Legendre multiplier sequences, since every Legendre multiplier sequence must also be a classical multiplier sequence (see [Blakeman et al. 2012, Theorem 8] together with [Piotrowski 2007, Proposition 118]). We note that if  $\{\gamma_k\}_{k=0}^\infty$  is a classical multiplier sequence, then one of

$$\{\gamma_k\}_{k=0}^\infty, \quad \{-\gamma_k\}_{k=0}^\infty, \quad \{(-1)^k \gamma_k\}_{k=0}^\infty, \quad \{(-1)^{k+1} \gamma_k\}_{k=0}^\infty$$

is a sequence of nonnegative terms [Pólya and Schur 1914, p. 90]. Since

$$\{-1\}_{k=0}^\infty \quad \text{and} \quad \{(-1)^k\}_{k=0}^\infty$$

are both classical multiplier sequences, it suffices to consider only sequences of nonnegative terms when characterizing classical multiplier sequences.

**Theorem 5** [Pólya and Schur 1914]. *Let  $\{\gamma_k\}_{k=0}^\infty$  be a sequence of nonnegative real numbers. The following are equivalent:*

- (1)  $\{\gamma_k\}_{k=0}^\infty$  is a classical multiplier sequence.
- (2) For each  $n$ , the polynomial  $T[(1+x)^n] := \sum_{k=0}^n \binom{n}{k} \gamma_k x^k$  is in  $\mathcal{L}\text{-}\mathcal{P}^+$ .
- (3)  $T[e^x] := \sum_{k=0}^\infty (\gamma_k/k!) x^k$  is in  $\mathcal{L}\text{-}\mathcal{P}^+$ .

Similar to the classical setting, we may consider only sequences of nonnegative terms when investigating (linear and cubic) Legendre multiplier sequences, by virtue of  $\{(-1)^k\}_{k=0}^\infty$  also being a Legendre multiplier sequence [Blakeman et al. 2012, Theorem 12].

We conclude this section by a theorem of Borcea and Brändén, which characterizes reality preserving linear operators  $T : \mathbb{R}[x] \rightarrow \mathbb{R}[x]$  in terms of their symbol

$G_T(x, y)$ . In order to be able to state their result (see Theorem 8), we need to make the following definitions.

**Definition 6.** The symbol of a linear operator  $T : \mathbb{R}[x] \rightarrow \mathbb{R}[x]$  is the formal power series defined by

$$G_T(x, y) := \sum_{n=0}^{\infty} \frac{(-1)^n T(x^n)}{n!} y^n.$$

**Definition 7.** A real polynomial  $p \in \mathbb{R}[x, y]$  is called stable if  $p(x, y) \neq 0$  whenever  $\text{Im}(x) > 0$  and  $\text{Im}(y) > 0$ . The Laguerre–Pólya class of real entire functions in two variables, denoted by  $\mathcal{L}\text{-}\mathcal{P}_2(\mathbb{R})$ , is the set of real entire functions in two variables, which are locally uniform limits in  $\mathbb{C}^2$  of real stable polynomials.

**Theorem 8** [Borcea and Brändén 2009]. A linear operator  $T : \mathbb{R}[x] \rightarrow \mathbb{R}[x]$  preserves the reality of zeros if and only if

- (1) the rank of  $T$  is at most 2 and  $T$  is of the form  $T(P) = \alpha(P)Q + \beta(P)R$ , where  $\alpha, \beta : \mathbb{R}[x] \rightarrow \mathbb{R}$  are linear functionals and  $Q + iR$  is a stable polynomial, or
- (2)  $G_T(x, y) \in \mathcal{L}\text{-}\mathcal{P}_2(\mathbb{R})$ , or
- (3)  $G_T(-x, y) \in \mathcal{L}\text{-}\mathcal{P}_2(\mathbb{R})$ .

In the remainder of this paper we follow the literature by using the notation  $T = \{\gamma_k\}_{k=0}^{\infty}$  to indicate the dual interpretation of a sequence as a linear operator and vice versa.

### 3. Linear Legendre sequences

We now reprove the nonexistence of linear Legendre multiplier sequences (see [Blakeman et al. 2012, Proposition 2]). Although the result is known, our proof is novel, and has the promise of being suitable for use when investigating  $Q$ -multiplier sequences in larger generality. The following definition and three lemmas serve as setup for Theorem 13.

**Definition 9.** We define a generalized hypergeometric function by

$${}_pF_q \left[ \begin{matrix} a_1, a_2, \dots, a_p \\ b_1, b_2, \dots, b_q \end{matrix} ; x \right] := 1 + \sum_{n=1}^{\infty} \frac{\prod_{i=1}^p (a_i)_n}{\prod_{j=1}^q (b_j)_n} \frac{x^n}{n!}, \tag{3-1}$$

where  $(\alpha)_n = \alpha(\alpha + 1) \cdots (\alpha + n - 1)$  denotes the rising factorial.

The convergence properties of the series on the right hand side of Equation (3-1) are discussed in detail in [Rainville 1960, Chapter 5]. Here we mention that if  $p = 3$  and  $q = 2$ , then the series is absolutely convergent on  $|x| = 1$  if

$$\Re \left( \sum_{j=1}^q b_j - \sum_{i=1}^p a_i \right) > 0.$$

**Lemma 10.** For all  $n \in \mathbb{N}^{\geq 1}$ , the generalized hypergeometric function

$${}_3F_2 \left[ \begin{matrix} -\frac{1}{2}, -n, \frac{1}{2} + n \\ \frac{1}{4}, \frac{3}{4} \end{matrix} ; -x \right]$$

converges at  $x = -1$  and satisfies the equation

$${}_3F_2 \left[ \begin{matrix} -\frac{1}{2}, -n, \frac{1}{2} + n \\ \frac{1}{4}, \frac{3}{4} \end{matrix} ; 1 \right] = 4n + 1.$$

*Proof.* Convergence at  $x = -1$  follows from the fact that

$$\Re\left(\frac{1}{4} + \frac{3}{4} - \left(-\frac{1}{2} - n + \frac{1}{2} + n\right)\right) = 1 > 0,$$

together with the remark after Definition 9. The rest of the claim follows directly from an application of Theorem 30 in [Rainville 1960], which states that for nonnegative integers  $n$ , and  $a, b$  independent of  $n$ , we have

$${}_3F_2 \left[ \begin{matrix} \frac{1}{2} + \frac{1}{2}a - b, -n, a + n \\ 1 + a - b, \frac{1}{2}a + \frac{1}{2} \end{matrix} ; 1 \right] = \frac{(b)_n}{(1 + a - b)_n}.$$

Setting  $a = \frac{1}{2}$  and  $b = \frac{5}{4}$  gives the required result. □

**Lemma 11.** Let  $n \in \mathbb{N}^{\geq 1}$ , and define

$$\Psi_n(x) := \sum_{j=1}^n \binom{n}{j} \frac{(2j-2)!}{(j-1)!} \frac{(\frac{1}{2} + 2j)_{n-j}}{(\frac{1}{2})_n} x^j.$$

Then

$${}_3F_2 \left[ \begin{matrix} -\frac{1}{2}, -n, \frac{1}{2} + n \\ \frac{1}{4}, \frac{3}{4} \end{matrix} ; -x \right] = 1 - 2\Psi_n(x).$$

*Proof.* The following identities are readily verified for  $0 \leq k \leq n$ .

$$(-1)^k \frac{(-n)_k}{k!} = \binom{n}{k}; \tag{3-2}$$

$$\left(\frac{1}{4}\right)_k \left(\frac{3}{4}\right)_k = \left(\frac{1}{2}\right)_{2k} 2^{-2k}; \tag{3-3}$$

$$2^k \left(-\frac{1}{2}\right)_k = -\frac{(2k-2)!}{2^{k-1}(k-1)!}; \tag{3-4}$$

$$\frac{(\frac{1}{2} + n)_k}{(\frac{1}{2})_{2k}} = \frac{(\frac{1}{2} + 2k)_{n-k}}{(\frac{1}{2})_n}. \tag{3-5}$$

With these in hand, we may now compute directly:

$$\begin{aligned}
 {}_3F_2 \left[ \begin{matrix} -\frac{1}{2}, -n, \frac{1}{2} + n \\ \frac{1}{4}, \frac{3}{4} \end{matrix} ; -x \right] &= \sum_{k=0}^{\infty} \frac{(-\frac{1}{2})_k (-n)_k (n + \frac{1}{2})_k}{(\frac{1}{4})_k (\frac{3}{4})_k k!} (-x)^k \\
 &= 1 + \sum_{k=1}^n (-1)^k \binom{n}{k} \frac{(n + \frac{1}{2})_k (-\frac{1}{2})_k}{(\frac{1}{2})_{2k} 2^{-2k}} (-x)^k \\
 &= 1 - \sum_{k=1}^n (-1)^k \binom{n}{k} \frac{(\frac{1}{2} + 2k)_{n-k} 2^k (2k - 2)!}{(\frac{1}{2})_n 2^{k-1} (k - 1)!} (-x)^k \\
 &= 1 - 2 \sum_{k=1}^n \binom{n}{k} \frac{(\frac{1}{2} + 2k)_{n-k} (2k - 2)!}{(\frac{1}{2})_n (k - 1)!} x^k \\
 &= 1 - 2\Psi_n(x),
 \end{aligned}$$

where the second equality uses equations (3-2) and (3-3), while the third equality employs equations (3-4) and (3-5). □

**Lemma 12.** *Let  $C_n := \binom{2n}{n} / (n + 1)$  denote the  $n$ -th Catalan number. For  $n \in \mathbb{N}^{\geq 1}$  the following equality holds:*

$$\begin{aligned}
 &\frac{C_{n-1}}{3 \cdot 2^{2n-2} (\frac{5}{2})_{2n-2}} \\
 &= \frac{1}{2^{2n} (\frac{1}{2})_{2n}} \left( 2n \frac{(-1)^n (\frac{1}{2})_n}{n!} + \sum_{j=1}^{n-1} \frac{C_{j-1}}{3 \cdot 2^{2j-2} (\frac{5}{2})_{2j-2}} \frac{(-1)^{n-j} (\frac{1}{2})_{n+j} 2^{2j}}{(n-j)!} \right).
 \end{aligned}$$

*Proof.* Note that the statement of the lemma is equivalent to

$$0 = 2n \frac{(-1)^n (\frac{1}{2})_n}{n!} + \sum_{j=1}^n \frac{C_{j-1} (-1)^{n-j} (\frac{1}{2})_{n+j}}{(\frac{1}{2})_{2j} (n-j)!} \quad \text{for all } n \in \mathbb{N}^{\geq 1}, \tag{3-6}$$

or

$$0 = 2n + \Psi_n(-1) \quad \text{for all } n \in \mathbb{N}^{\geq 1}, \tag{3-7}$$

where  $\Psi_n(x)$  is as in Lemma 11. Combining the results of Lemmas 10 and 11 gives

$$1 - 2\Psi_n(-1) = 4n + 1 \quad \text{for all } n \in \mathbb{N}^{\geq 1},$$

or equivalently,  $\Psi_n(-1) = -2n$  for  $n \geq 1$ . The proof is complete. □

We now prove the main theorem of the section.

**Theorem 13.** *Consider the operator  $T : \mathbb{R}[x] \rightarrow \mathbb{R}[x]$  given by*

$$T[\mathfrak{L}\mathfrak{e}_k(x)] = (k + c)\mathfrak{L}\mathfrak{e}_k(x) \quad \text{for } k = 0, 1, 2, 3, \dots \text{ and } c \in \mathbb{R}.$$

If we write  $T = \sum_{k=0}^{\infty} T_k(x)D^k$ , then

$$T_k(0) = \begin{cases} 0 & \text{if } k \text{ is odd,} \\ c & \text{if } k = 0, \\ -\frac{C_{n-1}}{3 \cdot 2^{2n-2} \left(\frac{5}{2}\right)_{2n-2}} & \text{if } k = 2n \text{ for } n \geq 1, \end{cases} \tag{3-8}$$

where  $C_n$  denotes the  $n$ -th Catalan number.

*Proof.* The following facts about Legendre polynomials are known explicitly, or follow easily from basic properties (see [Rainville 1960, pp. 157–158]):

(i) 
$$\mathfrak{L}e_n(x) = \frac{2^n \left(\frac{1}{2}\right)_n x^n}{n!} + \pi_{n-2} \quad (n \geq 0),$$

where  $\pi_{n-2}$  is a polynomial of degree  $n - 2$  in  $x$ .

(ii) 
$$\mathfrak{L}e_{2n+1}(0) = 0, \text{ for } n \geq 0.$$

(iii) 
$$\mathfrak{L}e_{2n}(0) = \frac{(-1)^n \left(\frac{1}{2}\right)_n}{n!} \quad (n \geq 0).$$

(iv) For  $0 \leq j \leq n$ ,

$$D^{2j} \mathfrak{L}e_{2n}(x) \Big|_{x=0} = \frac{(-1)^{n-j} \left(\frac{1}{2}\right)_{n+j} 2^{2j}}{(n-j)!},$$

while

$$D^{2j} \mathfrak{L}e_{2n+1}(x) \Big|_{x=0} = 0 \quad \text{for all } j, n \geq 0,$$

because Legendre polynomials with odd index are odd.

Mutatis mutandis, the proof of Proposition 29 in [Piotrowski 2007] demonstrates that the coefficient polynomials  $T_k(x)$  of the linear operator given in Theorem 13 can be computed recursively as

$$T_0(x) = T[1],$$

$$T_k(x) = \frac{1}{2^k \left(\frac{1}{2}\right)_k} \left( T[\mathfrak{L}e_k(x)] - \sum_{j=0}^{k-1} T_j(x)D^j[\mathfrak{L}e_k(x)] \right) \quad \text{for } k = 1, 2, 3, \dots$$

It is now easy to verify that  $T_0(x) = c$ ,  $T_1(x) = x$  and  $T_2(x) = -\frac{1}{3}$ , and the proposed values of  $T_k(0)$  follow readily for  $k = 0, 1, 2$ . Proceeding by induction we assume that  $T_j(0)$  is given by Equation (3-8) for  $0 \leq j \leq k - 1$  for some  $k \geq 1$ . If  $k$  is odd,

the second part of fact (iv) above yields

$$\begin{aligned} T_k(0) &= \frac{1}{2^k \left(\frac{1}{2}\right)_k} \left[ (k+c)\mathfrak{L}\mathfrak{e}_k(0) - \sum_{j=0}^{k-1} T_j(x) D^j [\mathfrak{L}\mathfrak{e}_k(x)] \Big|_{x=0} \right] \\ &= \frac{1}{2^k \left(\frac{1}{2}\right)_k} \left[ - \sum_{j=0}^{(k-1)/2} T_{2j}(x) D^{2j} [\mathfrak{L}\mathfrak{e}_k(x)] \Big|_{x=0} \right] \\ &= 0. \end{aligned}$$

On the other hand, if  $k$  is even, writing  $k = 2n$  and using the first part of fact (iv) gives

$$\begin{aligned} T_k(0) &= \frac{1}{2^{2n} \left(\frac{1}{2}\right)_{2n}} \left[ (2n+c) \frac{(-1)^n \left(\frac{1}{2}\right)_n}{n!} - \sum_{j=0}^{k-1} T_j(x) D^j [\mathfrak{L}\mathfrak{e}_k(x)] \Big|_{x=0} \right] \\ &= \frac{1}{2^{2n} \left(\frac{1}{2}\right)_{2n}} \left[ 2n \frac{(-1)^n \left(\frac{1}{2}\right)_n}{n!} - \sum_{j=1}^{k-1} T_j(x) D^j [\mathfrak{L}\mathfrak{e}_k(x)] \Big|_{x=0} \right] \\ &= \frac{1}{2^{2n} \left(\frac{1}{2}\right)_{2n}} \left[ 2n \frac{(-1)^n \left(\frac{1}{2}\right)_n}{n!} - \sum_{j=1}^{(k-2)/2} T_{2j}(x) D^{2j} [\mathfrak{L}\mathfrak{e}_k(x)] \Big|_{x=0} \right] \\ &= \frac{1}{2^{2n} \left(\frac{1}{2}\right)_{2n}} \left[ 2n \frac{(-1)^n \left(\frac{1}{2}\right)_n}{n!} + \sum_{j=1}^{n-1} \frac{C_{j-1}}{3 \cdot 2^{2j-2} \left(\frac{5}{2}\right)_{2j-2}} \frac{(-1)^{n-j} \left(\frac{1}{2}\right)_{n+j} 2^{2j}}{(n-j)!} \right] \\ &= - \frac{C_{n-1}}{3 \cdot 2^{2n-2} \left(\frac{5}{2}\right)_{2n-2}}, \end{aligned}$$

where the last equality is the result of Lemma 12. □

Let  $T$  be the operator corresponding to the Legendre sequence  $\{k+c\}_{k=0}^\infty$ . Recall that the symbol of  $T$  is given by

$$G_T(-x, y) = \sum_{k=0}^\infty \frac{(-1)^k T[x^k] y^k}{k!},$$

and that  $T$  is reality preserving (that is,  $\{k+c\}_{k=0}^\infty$  is a Legendre multiplier sequence) if and only if either  $G_T(-x, y)$  or  $G_T(x, y)$  belongs to  $\mathcal{L}\text{-}\mathcal{P}_2(\mathbb{R})$ , since the sequence under consideration is nontrivial. Following [Brändén and Ottergren 2014], we expand  $G_T(-x, y)$  and  $G_T(x, y)$  as a series in powers of  $x$ . By Theorem 13 the constant term in both of these expansions is

$$f(y) := c - \sum_{k=1}^\infty \frac{C_{k-1} y^{2k}}{3 \cdot 2^{2k-2} \left(\frac{5}{2}\right)_{2k-2}}.$$

Thus  $f(y) \in \mathcal{L}\text{-}\mathcal{P}$  if either  $G_T(-x, y)$  or  $G_T(x, y)$  were in  $\mathcal{L}\text{-}\mathcal{P}_2(\mathbb{R})$ , since we obtain  $f(y)$  from either  $G_T(-x, y)$  or  $G_T(x, y)$  by applying the nonnegative multiplier sequence  $1, 0, 0, 0, \dots$  acting on  $x$ , which preserves the class  $\mathcal{L}\text{-}\mathcal{P}_2(\mathbb{R})$  (see [Borcea and Brändén 2010; Brändén 2014]). We shall now demonstrate that  $f(y)$  is an entire function which does not belong to the Laguerre–Pólya class, and hence  $\{k + c\}_{k=0}^\infty$  is not a Legendre multiplier sequence for any  $c \in \mathbb{R}$ .

**Proposition 14.** *Let  $c \in \mathbb{R}$ . Then*

$$f(y) = c - \sum_{k=1}^\infty \frac{C_{k-1}y^{2k}}{3 \cdot 2^{2k-2} \left(\frac{5}{2}\right)_{2k-2}}$$

is an entire function which does not belong to  $\mathcal{L}\text{-}\mathcal{P}$ .

*Proof.* Consider the change of variables  $x = y^2$  and the function

$$\tilde{f}(x) = c - \frac{4}{3} \sum_{k=1}^\infty \frac{C_{k-1}x^k}{2^{2k} \left(\frac{5}{2}\right)_{2k-2}} = c - \frac{4}{3} \sum_{k=1}^\infty a_k x^k.$$

Since

$$(\star) \quad \lim_{k \rightarrow \infty} \frac{a_{k+1}}{a_k} = \lim_{k \rightarrow \infty} \frac{2(2k-1)}{k+1} \frac{1}{(5+2(2k-2))(5+2(2k-1))} = 0,$$

$\tilde{f}(x)$  is entire. The existence of the limit in  $(\star)$  implies that  $\lim_{k \rightarrow \infty} \sqrt[k]{a_k} = 0$  as well, and hence  $f(y)$  is also entire.

It remains to show that  $f(y) \notin \mathcal{L}\text{-}\mathcal{P}$ . To this end, we first demonstrate that  $\tilde{f}(x) \notin \mathcal{L}\text{-}\mathcal{P}$ . Writing  $d_k = k! a_k$  we can express  $\tilde{f}(x)$  as

$$\tilde{f}(x) = c - \frac{4}{3} \sum_{k=1}^\infty \frac{d_k}{k!} x^k.$$

By Theorem 4 and the comments thereafter, if  $\tilde{f}(x)$  were to belong to  $\mathcal{L}\text{-}\mathcal{P}$ , we would have  $L_1(x, \tilde{f}^{(k)}) \geq 0$  for all  $k = 0, 1, 2, \dots$  and  $x \in \mathbb{R}$ . In particular,  $L_1(0, \tilde{f}') = \frac{16}{9}(d_2^2 - d_3d_1) \geq 0$  would hold. A quick calculation reveals that

$$d_2^2 - d_3d_1 = -\frac{1}{80850} < 0,$$

establishing that  $\tilde{f}(x) \notin \mathcal{L}\text{-}\mathcal{P}$ . Suppose now that  $f(y) \in \mathcal{L}\text{-}\mathcal{P}$ . By virtue of being an even function,  $f(y)$  has the factorization

$$f(y) = ce^{-ay^2} \prod_{k=1}^\omega \left(1 - \frac{y^2}{x_k^2}\right),$$

where  $a \geq 0$ ,  $x_k \in \mathbb{R} \setminus \{0\}$ ,  $0 \leq \omega \leq \infty$ , and  $\sum 1/x_k^2 < \infty$ . Replacing  $y^2$  by  $x$  would yield  $\tilde{f}(x) \in \mathcal{L}\text{-}\mathcal{P}$ , a contradiction. We conclude that  $f(y) \notin \mathcal{L}\text{-}\mathcal{P}$ , and our proof is complete.  $\square$

#### 4. Cubic Legendre multiplier sequences

In this section we establish the nonexistence of cubic Legendre multiplier sequences. Without loss of generality we may consider sequences interpolated by monic polynomials. Since every such cubic polynomial can be written as  $(k^2 + \alpha k + \beta)(k + c)$  for some real triple  $(\alpha, \beta, c)$ , one may wish to proceed based on whether or not the quadratic factor in the product is itself a Legendre multiplier sequence. It turns out that such case analysis is more than one needs: we can handle all cubic sequences at once. We begin with two preparatory results.

**Lemma 15.** *Suppose  $T = \{k^3 + ak^2 + bk + c\}_{k=0}^\infty$  is a sequence of nonnegative terms. If  $T$  is a classical multiplier sequence, then  $a \geq -3$ ,  $a + b \geq -1$  and  $c \geq 0$ .*

*Proof.* By Theorem 5,  $T$  is a classical multiplier sequence if and only if  $T[e^x] \in \mathcal{L}\text{-}\mathcal{P}^+$ . We have

$$\begin{aligned} T[e^x] &= \sum_{k=0}^\infty (k^3 + ak^2 + bk + c) \frac{x^k}{k!} \\ &= e^x (x^3 + (a + 3)x^2 + (a + b + 1)x + c). \end{aligned}$$

Thus the coefficients of the polynomial

$$x^3 + (a + 3)x^2 + (a + b + 1)x + c$$

must all be nonnegative. The claim follows.  $\square$

**Lemma 16** [Levin 1956, Lemma 3, p. 337]. *If all zeros of the real polynomial*

$$h(x) = c_0 + c_1x + \dots + c_nx^n \quad (c_n \neq 0)$$

*are real,  $c_0 \neq 0$  and  $c_p = 0$  for some  $0 < p < n$ , then  $c_{p-1}c_{p+1} < 0$ .*

We are now ready to state and prove the main theorem of the section.

**Theorem 17.** *The sequence  $\{k^3 + ak^2 + bk + c\}_{k=0}^\infty$  is not a Legendre multiplier sequence for any real triple  $(a, b, c)$ .*

*Proof.* Denote by  $T_{a,b,c}$  the operator associated to the Legendre sequence

$$\{k^3 + ak^2 + bk + c\}_{k=0}^\infty.$$

By Lemma 15, in order for  $\{k^3 + ak^2 + bk + c\}_{k=0}^\infty$  to be a classical multiplier sequence we must have  $a \geq -3$ ,  $a + b \geq -1$  and  $c \geq 0$ . Consider now the action



of  $T_{\alpha,\beta,c}$  on the two polynomials

$$\begin{aligned} p_1(x) &= x^5 \mathfrak{L}e_3(x) \\ &= \frac{64}{1287} \mathfrak{L}e_8(x) + \frac{152}{693} \mathfrak{L}e_6(x) + \frac{372}{1001} \mathfrak{L}e_4(x) + \frac{205}{693} \mathfrak{L}e_2(x) + \frac{4}{63}, \\ p_2(x) &= x^5 \mathfrak{L}e_5(x) \\ &= \frac{2016}{46189} \mathfrak{L}e_{10}(x) + \frac{4816}{24453} \mathfrak{L}e_8(x) + \frac{4078}{11781} \mathfrak{L}e_6(x) \\ &\quad + \frac{291}{1001} \mathfrak{L}e_4(x) + \frac{1000}{9009} \mathfrak{L}e_2(x) + \frac{8}{693}. \end{aligned}$$

Computing  $18018T_{a,b,c}[p_1(x)] = \sum_{k=0}^4 q_{2k}(a, b, c)x^{2k}$ , we find that

$$\begin{aligned} q_0(a, b, c) &= 16(-121 + 46a - 46b), \\ q_4(a, b, c) &= 630(15724 + 1226a + 61b), \end{aligned}$$

with the restrictions on  $a, b$  and  $c$  implying directly that  $q_4(a, b, c) > 0$  for all real triples  $(a, b, c)$  under consideration. If  $q_2(a, b, c) = 0$ , then reversing coefficients, and taking four derivatives of  $T_{a,b,c}[p_1(x)]$  (both of which operations preserve the reality of zeros) results in a polynomial with nonreal zeros. If  $q_2(a, b, c) \neq 0$ , then in light of Lemma 16, a necessary condition for  $T_{a,b,c}[p_1(x)]$  to have only real zeros is that

$$(\dagger) \quad q_0(a, b, c) = 16(-121 + 46a - 46b) \geq 0.$$

We now turn our attention to  $T_{a,b,c}[p_2(x)]$ . If we write  $23279256T_{a,b,c}[p_2(x)] = \sum_{k=0}^5 w_{2k}(a, b, c)x^{2k}$ , then

$$\begin{aligned} w_0(a, b, c) &= 16(-641 + 806a - 806b), \\ w_4(a, b, c) &= -630(38840980 + 2015774a + 62731b), \end{aligned}$$

with Lemma 15 implying that  $w_4(a, b, c) < 0$  for all admissible triples  $(a, b, c)$ . Considerations identical to those above imply that either  $T_{a,b,c}[p_2(x)]$  has nonreal zeros, or the inequality

$$(\ddagger) \quad w_0(a, b, c) = 16(-641 + 806a - 806b) \leq 0$$

must hold. Combining inequalities  $(\dagger)$  and  $(\ddagger)$  we obtain

$$-\frac{121}{46} + a \geq b \geq -\frac{641}{806} + a,$$

a clear impossibility. We conclude that  $T_{a,b,c}$  cannot simultaneously preserve the reality of the zeros of  $x^5 \mathfrak{L}e_3(x)$  and  $x^5 \mathfrak{L}e_5(x)$ . Whence  $\{k^3 + ak^2 + bk + c\}_{k=0}^\infty$  is not a Legendre multiplier sequence for any real triple  $(a, b, c)$ .  $\square$

**Remark 18.** Theorem 17 yields yet another proof of the nonexistence of linear Legendre multiplier sequences by the following considerations. If  $T_1, T_2$  are Legendre

multiplier sequences, then so is  $T_1T_2$ . Since  $\{k^2 + k + \beta\}$  is a Legendre multiplier sequence whenever  $\beta \in [0, 1]$ , the existence of linear Legendre multiplier sequences would immediately imply the existence of cubic Legendre multiplier sequences, contradicting Theorem 17.

### 5. Open problems

The following is a list of open problems motivated by the preceding results. These questions are not only related to the classification of Legendre multiplier sequences but also to some general properties of reality preserving linear operators  $T = \sum_{k=0}^{\infty} T_k(x)D^k$  on  $\mathbb{R}[x]$ , properties which are captured in the coefficient polynomials  $T_k(x)$ .

**5.1. Higher order Legendre sequences.** The characterization of polynomials with degree four or higher which interpolate Legendre multiplier sequences remains open. Using computational techniques as in Section 4 quickly turns intractable with the increasing number of parameters. In addition, one has to judiciously select “test polynomials” in order for this method to succeed succinctly. The polynomials

$$p(n, k) = x^k \mathcal{L}e_n(x)$$

mimic properties of the test polynomials  $(1+x)^n$  for classical multiplier sequences in that they have zeros of high multiplicity away from the zeros of the basis polynomials. As such, we were able to use just a couple test polynomials to demonstrate the nonexistence of cubic Legendre multiplier sequences. On the downside, the degrees of these polynomials are high and we believe that the degrees of the test polynomials would have to increase if one would want to eliminate sequences interpolated by higher order polynomials.

**5.2. Monotone operators.** We call an operator  $T = \sum_{k=0}^{\infty} T_k(x)D^k$  monotone if  $\deg T_k(x) \geq \deg T_{k-1}(x)$  for all  $k = 1, 2, \dots$ . The operator corresponding to the linear Legendre sequence  $\{k + c\}_{k=0}^{\infty}$  is given by

$$T = c + xD - \frac{1}{3}D^2 + \frac{2}{15}xD^3 + \sum_{k=4}^{\infty} T_k(x)D^k,$$

whereas the operator corresponding to the Legendre sequence  $\{k^2 + \alpha k + \beta\}_{k=0}^{\infty}$ ,  $\alpha \neq 1$ , is given by

$$T = \beta + (1 + \alpha)xD - \frac{2 + \alpha - 3x^2}{3}D^2 + \frac{2}{15}(\alpha - 1)x D^3 - \frac{(\alpha - 1)(1 + 4x^2)}{105}D^4 + (\alpha - 1) \sum_{k=5}^{\infty} T_k(x)D^k.$$

Neither sequence is a Legendre multiplier sequence, and neither operator is monotone. We believe these facts to be related, and give the following

**Conjecture 19.** *Suppose  $T = \sum_{k=0}^{\infty} T_k(x)D^k$  is an infinite order differential operator. If  $T$  is not monotone, then  $T$  is not reality preserving.*

Should this conjecture be true, one could then try to prove that if

$$\{\gamma_k\}_{k=0}^{\infty} = \{p(k)\}_{k=0}^{\infty},$$

where  $\deg p$  is odd and  $\{\gamma_k\}_{k=0}^{\infty}$  is a Legendre sequence, the operator corresponding to the sequence  $\{\gamma_k\}_{k=0}^{\infty}$  is an infinite order differential operator which is not monotone.

**5.3. Using the symbol of the operator.** Our approach used in Section 3 could be extended to treat sequences interpolated by higher order polynomials. Piotrowski [2007] and Forgács and Piotrowski [2013a] give explicit representations of the coefficient polynomials  $T_k(x)$  of classical, and Hermite diagonal operators respectively. In both cases the  $T_k(x)$ s are given in terms of the reverses of the Jensen polynomials associated to the sequence  $\{\gamma_k\}_{k=0}^{\infty}$ . If a sequence  $\{\gamma_k\}_{k=0}^{\infty}$  is interpolated by a polynomial, then only finitely many of these reverse Jensen polynomials are nonzero. This means that an analog of Theorem 13 would need the identification of only finitely many sequences, one for each reverse Jensen polynomial involved in the  $T_k(x)$ s, in order to explicitly determine the sequence  $\{T_k(0)\}_{k=0}^{\infty}$ . With this sequence in hand, one could carry out steps analogous to those in Section 3 to establish the nonexistence of Legendre multiplier sequences interpolated by polynomials of degree greater than three.

### Acknowledgements

We would like to thank George Csordas for many stimulating discussions and guiding insights, and the anonymous referee for numerous suggestions improving the exposition and streamlining of proofs in Lemma 11 and Theorem 17.

### References

- [Blakeman et al. 2012] K. Blakeman, E. Davis, T. Forgács, and K. Urabe, “On Legendre multiplier sequences”, *Missouri J. Math. Sci.* **24**:1 (2012), 7–23. MR 2977127
- [Borcea and Brändén 2009] J. Borcea and P. Brändén, “Pólya–Schur master theorems for circular domains and their boundaries”, *Ann. of Math. (2)* **170**:1 (2009), 465–492. MR 2010g:30004 Zbl 1184.30004
- [Borcea and Brändén 2010] J. Borcea and P. Brändén, “Multivariate Pólya–Schur classification problems in the Weyl algebra”, *Proc. Lond. Math. Soc. (3)* **101**:1 (2010), 73–104. MR 2012a:47075 Zbl 1196.47028

- [Brändén 2014] P. Brändén, “The Lee–Yang and Pólya–Schur programs, III: Zero-preservers on Bargmann–Fock spaces”, *Amer. J. Math* **136**:1 (2014), 241–253. arXiv 1107.1809
- [Brändén and Ottergren 2014] P. Brändén and E. Ottergren, “A characterization of multiplier sequences for generalized Laguerre bases”, *Constr. Approx.* **39**:3 (2014), 585–596. MR 3207673
- [Csordas and Varga 1990] G. Csordas and R. S. Varga, “Necessary and sufficient conditions and the Riemann hypothesis”, *Adv. in Appl. Math.* **11**:3 (1990), 328–357. MR 91d:11107 Zbl 0707.11062
- [Csordas and Vishnyakova 2013] G. Csordas and A. Vishnyakova, “The generalized Laguerre inequalities and functions in the Laguerre–Pólya class”, *Cent. Eur. J. Math.* **11**:9 (2013), 1643–1650. MR 3071931 Zbl 06236721
- [Forgács and Piotrowski 2013a] T. Forgács and A. Piotrowski, “Hermite multiplier sequences and their associated operators”, preprint, 2013. To appear in *Constr. Approx.* arXiv 1312.6187
- [Forgács and Piotrowski 2013b] T. Forgács and A. Piotrowski, “Multiplier sequences for generalized Laguerre bases”, *Rocky Mountain J. Math.* **43**:4 (2013), 1141–1159. MR 3105315 Zbl 06209831
- [Levin 1956] B. J. Levin, *Распределение корней целых функций*, State Publishing House of Technical and Theoretical Literature, Moscow, 1956. Translated as *Distribution of zeros of entire functions*, Transl. Math. Monogr. **5**, American Mathematical Society, Providence, RI, 1964, revised ed. in 1980. MR 28 #217 Zbl 0111.07401
- [Obreschkoff 1963] N. Obreschkoff, *Verteilung und Berechnung der Nullstellen reeller Polynome*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1963. MR 29 #1302 Zbl 0156.28202
- [Piotrowski 2007] A. Piotrowski, *Linear operators and the distribution of zeros of entire functions*, thesis, University of Hawaii, Manoa, HI, 2007, available at <http://hdl.handle.net/10125/25932>. MR 2710244
- [Pólya and Schur 1914] G. Pólya and J. Schur, “Über zwei Arten von Faktorenfolgen in der Theorie der algebraischen Gleichungen”, *J. Reine Angew. Math.* **144** (1914), 89–113. JFM 45.0176.01
- [Rainville 1960] E. D. Rainville, *Special functions*, Macmillan, New York, 1960. MR 21 #6447 Zbl 0092.06503
- [Yoshida 2013] R. Yoshida, *Linear and non-linear operators, and the distribution of zeros of entire functions*, Ph.D. thesis, University of Hawaii, Manoa, HI, 2013, available at <http://hdl.handle.net/10125/29455>.

Received: 2013-10-16

Revised: 2014-01-09

Accepted: 2014-01-24

tforgacs@csufresno.edu

*Department of Mathematics, California State University  
Fresno, 5245 North Backer Ave, M/S, PB 108,  
Fresno, CA 93740-8001, United States*

jhaley5@u.rochester.edu

*Mathematics, University of Rochester, 915 Hylan Building,  
RC Box 270138, Rochester, NY 14627, United States*

ram0022@tigermail.auburn.edu

*Department of Mathematics and Statistics, Auburn University,  
221 Parker Hall, Auburn, AL 36849, United States*

casimon@davidson.edu

*Department of Mathematics, Davidson College, Box 7129,  
Davidson, NC 28035, United States*

# Seating rearrangements on arbitrary graphs

Daryl DeFord

(Communicated by Kenneth S. Berenhaut)

We exhibit a combinatorial model based on seating rearrangements, motivated by some problems proposed in the 1990s by Kennedy, Cooper, and Honsberger. We provide a simpler interpretation of their results on rectangular grids, and then generalize the model to arbitrary graphs. This generalization allows us to pose a variety of well-motivated counting problems on other frequently studied families of graphs.

## 1. Introduction

**1.1. Background.** In this section we describe the original motivation for our problems and the original interpretations that are present in the literature.

**1.1.1. Original problem.** Our interest in this combinatorial model begins with a problem presented by Honsberger [1997]:

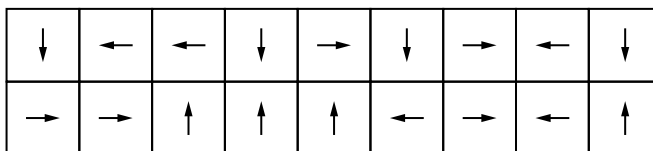
A classroom has 5 rows of 5 desks per row. The teacher requests each pupil to change his seat by going either to the seat in front, the one behind, the one to his left, or the one on his right (of course not all these options are possible for all students). Determine whether or not this directive can be carried out.

It can easily be shown that this directive is impossible [Honsberger 1997; Kennedy and Cooper 1993]. Consider coloring the classroom like a checkerboard. Then every student initially placed on a “white desk” must move to a “black desk” and vice versa. However, our chessboard coloring has 13 white squares and 12 black squares. Thus, were such a rearrangement to exist, by the pigeonhole principle there must be at least one black desk that receives two students from white squares and this violates the terms of the directive. More generally, this proof obviously generalizes to any rectangular classroom that has both an odd number of rows and columns [Otake et al. 1996].

---

*MSC2010:* 05C30.

*Keywords:* matrix permanents, cycle covers, tilings, recurrence relations.



**Figure 1.** A  $2 \times 9$  seating rearrangement.

**1.1.2. Early work.** Curtis Cooper and Robert Kennedy [1993] explored some basic extensions to this rearrangement problem by applying some traditional combinatorial and linear algebraic techniques (see also [Otake et al. 1996]). Their goal was to solve the following more general problem:

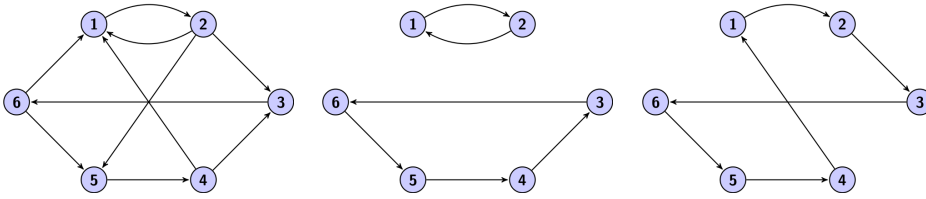
A classroom has  $m$  rows of  $n$  desks per row. The teacher requests each pupil to change his seat by going either to the seat in front, the one behind, the one to his left, or the one on his right (of course not all these options are possible for all students). *In how many ways* can this directive be carried out?

They began by solving the  $2 \times n$  and  $3 \times n$  cases by classifying all possible endings and constructing matrix systems that represented the interactions among these endings. For example, Figure 1 shows a  $2 \times 9$  seating rearrangement. Then, the principle of mathematical induction can be used to show that the constructed matrix systems faithfully represent the counting problem. Of particular interest is the fact that the number of rearrangements on a  $2 \times n$  grid is equal to the square of the  $(n + 1)$ -st Fibonacci number. In Section 2.1 we will give a combinatorial proof of this fact. However, this method quickly becomes unwieldy, and they were forced to seek more powerful tools to solve the general case.

In order to count the  $2m \times n$  seating rearrangements, Cooper and Kennedy turned to the theory of matrix permanents [Marcus and Minc 1965; Otake et al. 1996]. By modifying the adjacency matrix of the underlying grid graph and taking a symbolic determinant of the resulting block matrix they obtained the following representation of the number of seating rearrangements of a  $2m \times n$  classroom:

$$2^{2mn} \prod_{t=1}^{2m} \prod_{s=1}^n \left( \cos^2 \left( \frac{s\pi}{n+1} \right) + \cos^2 \left( \frac{t\pi}{2m+1} \right) \right). \quad (1-1)$$

This formula is very similar to the expression derived in 1961 by Kasteleyn [Harary 1967; Kasteleyn 1961], and Temperley and Fisher [1961], that counts the number of domino tilings of a  $m \times n$  grid. In Section 2.1 we will justify this correspondence while in Section 4 we will prove a general theorem that gives this relationship as an immediate corollary.



**Figure 2.** Two cycle covers of a digraph. Left: digraph; middle: even cover; right: odd cover.

**1.2. Mathematical preliminaries.** The proofs and results in this paper rely on techniques from combinatorics, linear algebra, and graph theory. Basic definitions and notation not presented here can be found in [Chartrand et al. 2011; van Lint and Wilson 2001; Shilov 1977].

**1.2.1. Cycle covers.** Given a digraph  $D = \{V, E\}$ , a cycle cover is defined as a subset of the edges,  $C \subseteq E$ , such that the induced digraph on  $C$  contains each vertex of  $V$  and each of those vertices lies on exactly one cycle [Harary 1969]. It is easy to see that each cycle cover of a digraph can be considered a permutation of the set of vertex labels, and more specifically a derangement, if no self-loops occur in the digraph. Thus, it is reasonable to consider the parity of a given cycle cover, defined as the parity of the permutation it represents.

Hence, a cycle cover that contains an even number of even cycles is considered even, while a cycle cover with an odd number of even cycles is considered odd. Figure 2 shows a digraph and two of its cycle covers, one of each parity.

**1.2.2. Matrix permanents.** The permanent of a matrix,  $M$ , with elements,  $M_{u,v}$ , is defined as the unsigned sum over all of the permutations of the matrix [Harary 1969; Marcus and Minc 1965]. Thus,

$$\text{per } M = \sum_{\pi \in S_n} \prod_{i=1}^n M_{i,\pi(i)}, \quad (1-2)$$

is a symbolic representation of the matrix permanent. It is computationally difficult to calculate the permanent of a general 0–1 matrix (technically the problem of computing the permanent is  $\#P$  complete) [Aaronson 2011; Lundow 1996; Valiant 1979]. Although the definition of the permanent looks very similar to that of the determinant, the permanent shares very few of the determinant's useful algebraic properties or relations to eigenvalues. Also, the determinant of a matrix can be calculated in polynomial time by Gaussian elimination, while the permanent cannot. However, interchanging rows or columns of the matrix does not affect the value of the permanent of that matrix [Marcus and Minc 1965].

We are interested in the concept of matrix permanents because the permanent of the adjacency matrix of a digraph is equal to the number of cycle covers of that digraph [Harary 1969]. A survey of results in combinatorics based on this method can be found in [Kuperberg 1998]. However, since the permanent is often infeasible to compute, a natural question is to ask whether we can change the signs of some elements of a given adjacency matrix,  $A$ , to form a new matrix,  $A'$ , with the property that:

$$\text{per } A = \det A'. \quad (1-3)$$

This question of “convertible” matrices was originally posed by Pólya [1913]. Beineke and Harary [1966] showed that digraphs whose adjacency matrix admits an orientation satisfying (1-3) are exactly those that contain no odd cycle covers. Later, Vazarani and Yannakakis [1988] proved that this problem is equivalent to finding pfaffian orientations of bipartite graphs. The pfaffian of a skew-symmetric matrix is a sum over signed products of entries in the matrix that can be used to count the number of perfect matchings in some graphs. For a complete discussion of pfaffians and their relation to perfect matchings see [Loehr 2011, Chapter 12.12].

This problem of pfaffians was characterized by Little [1975], who showed that a given bipartite graph,  $B$ , admits a pfaffian orientation if and only if  $B$  contains no subgraph homeomorphic to  $K_{3,3}$  (the complete bipartite graph with three vertices in each partite set). An obvious extension of this question is to ask how difficult it is to construct such a matrix  $A'$  given  $A$ . Finally Robertson, Seymour, and Thomas [Robertson et al. 1999] settled the issue by giving a polynomial time algorithm that takes a given graph and either constructs an orientation of its adjacency matrix that satisfies (1-3), or demonstrates a subgraph of  $G$  proving that (1-3) cannot be satisfied.

## 2. Seating rearrangements

In this section we motivate and present our basic model through some simple counting problems.

**2.1. Domino tilings.** The original problem studied by Cooper and Kennedy can easily be expressed in terms of perfect matchings or domino tilings, both of which are very familiar combinatorial objects. We showed previously that if  $m$  and  $n$  are both odd there can be no legitimate rearrangements in an  $m \times n$  classroom, so we will only consider the cases where at least one of  $m$  and  $n$  are even. However, note that the case where there are no legitimate rearrangements trivially satisfies the following lemma as there are no perfect matchings on  $P_m \times P_n$  when  $m$  and  $n$  are both odd, where  $P_k$  is the path graph on  $k$  vertices.



**Lemma 1.** *The number of legitimate seating rearrangements in a  $2m \times n$  classroom is equal to the square of the number of domino tilings of a  $2m \times n$  grid.*

*Proof.* Begin by coloring the classroom like a chessboard. Note that we may consider the rearrangements of the students initially sitting in white desks separately from the rearrangements of those sitting in black desks since the two groups cannot interfere with each other. Since there are exactly as many black desks as white desks, arranged in the same fashion, the total number of rearrangements is equal to the square of the number of either the black or white rearrangements computed separately.

To complete the proof, consider tiling a  $2m \times n$  board with  $mn$  dominoes. We can construct a bijection between the rearrangements of students initially placed in black (white) desks with domino tilings by placing a domino in the tiling for each student that covers that student's initial desk and their destination desk. Thus, any seating rearrangement can be deconstructed into two independent domino tilings, one for each initial color. Figure 3 gives an example of this process.

In order to construct a seating rearrangement from an independently selected pair of domino tilings we may perform the operation in reverse. Without loss of generality, associate one of the tilings with movements from white desks to black desks, and associate the other tiling with movements from black desks to white desks. Hence, we can combine any two domino tilings to create a unique seating rearrangement and the proof is complete.  $\square$

It is well known (and can be easily seen by comparison to  $1 \times n$  tilings with squares and dominoes), that the number of domino tilings of a  $2 \times n$  rectangle is equal to the  $(n + 1)$ -st Fibonacci number. This observation, combined with the preceding lemma, provides a combinatorial explanation for the inductive-matrix result of Cooper and Kennedy mentioned in the introduction:

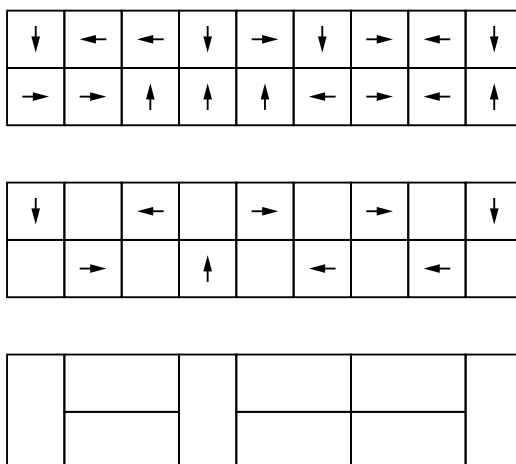
**Corollary 2.** *The number of seating rearrangements on a  $2 \times n$  classroom is equal to the square of the  $(n + 1)$ -st Fibonacci number.*

Another natural corollary to this lemma is a special case of Theorem 8, which will be proved in Section 4.

**Corollary 3.** *The number of legitimate seating rearrangements in an  $m \times 2n$  classroom is equal to the square of the number of perfect matchings on  $P_{2m} \times P_n$ .*

**2.2. Arbitrary graphs.** In order to extend this notion of seating rearrangements to arbitrary graphs we constructed the following modified problem statement:

**Problem.** Given a graph, place a marker on each vertex. We want to count the number of legitimate "rearrangements" of these markers subject to the following rules:



**Figure 3.** A rearrangement/tiling correspondence.

- Each marker must move to an adjacent vertex.
- After all of the markers have moved, each vertex must contain exactly one marker.

Thus, we define the number of rearrangements on an arbitrary graph to be the number of ways to satisfy the requirements given above. A related, interesting problem is to consider rearrangements where the markers are allowed to *either* remain in place or move along an edge to an adjacent vertex. To formulate this problem extension in graph-theoretic terms, we can add a self-loop to each vertex in the graph and proceed with the problem statement given above, where a vertex with a self-loop is considered adjacent to itself.

**2.3. Digraphs.** Given any graph  $G$ , we can construct a digraph  $\vec{G}$ , by replacing each simple edge of  $G$  by a pair of directed edges, one in each orientation. Then, the following lemma shows that there is a one-to-one correspondence between rearrangements on  $G$  and cycle covers on  $\vec{G}$ .

**Lemma 4.** *The number of rearrangements on any simple graph  $G$  is equal to the number of cycle covers on  $\vec{G}$ .*

*Proof.* Consider a legitimate rearrangement on a graph  $G$ , under the rules presented above. To construct a unique cycle cover on  $\vec{G}$ , place a directed edge in the cycle cover beginning at each markers initial vertex and ending at that markers terminal vertex. By the first rule, each vertex must have out-degree equal to 1. Similarly, by the second rule, each vertex must have in-degree equal to 1. Hence, the constructed cycle cover spans all vertices of  $G$  and has  $d^+(v) = d^-(v) = 1$  for all  $v \in V(G)$ , and so is a legitimate cycle cover.

A unique rearrangement on  $G$  can be constructed from a given cycle cover on  $\vec{G}$  in a similar fashion. Thus, there exists a bijection between these rearrangements and cycle covers, which implies that their magnitudes are equal.  $\square$

This gives us the following method for counting rearrangements on arbitrary graphs as well as a combinatorial interpretation of a matrix permanent of the adjacency matrix of a simple graph.

**Lemma 5.** *Given a graph  $G$ , with adjacency matrix  $A(G)$ , the number of rearrangements on  $G$  is equal to  $\text{per}(A(G))$ .*

*Proof.* By construction, the adjacency matrices of  $G$  and  $\vec{G}$  are equal, and the permanent of the adjacency matrix of  $\vec{G}$  is equal to the number of cycle covers on  $\vec{G}$ . Since, by Lemma 4, there is a one-to-one correspondence between cycle covers on  $\vec{G}$  and legitimate rearrangements on  $G$ , this proof is complete.  $\square$

Hence, we have a numerical method to compute the number of rearrangements on any graph. This method is computationally inefficient in general, but can provide numerical values of initial conditions for recurrence relations and generating functions, as well as providing empirical evidence of growth rates and divisibility properties.

**2.4. Notation.** For the rest of this paper we will use the notation  $R(G)$  to represent the number of legitimate rearrangements on a given graph  $G$ . Similarly,  $R_s(G)$  will represent the number of rearrangements where each marker is allowed to remain in place. Thus, the statement in the previous lemma could be rewritten as  $R(G) = \text{per}(A(G))$ .

Several times throughout this paper, we will use the Fibonacci numbers in our counting. In these instances we will use the combinatorial Fibonacci numbers  $f_n = F_{n+1}$ , indexed as  $f_0 = 1$ ,  $f_1 = 1$  and  $f_n = f_{n-1} + f_{n-2}$  for  $n \geq 2$ . This indexing is motivated by the traditional counting interpretation of the Fibonacci numbers as the number of ways to tile a  $1 \times n$  board with squares and dominoes. Similarly, we will also employ the Lucas numbers,  $l_n$ , with  $l_n = f_n + f_{n-2}$  defined as the number of ways to tile a  $1 \times n$  bracelet with “rounded” squares and dominoes [Benjamin and Quinn 2003].

From graph theory,  $K_n$  will represent the complete graph on  $n$  vertices, while  $K_{m,n}$  will be the complete bipartite graph with bipartite sets of order  $m$  and  $n$ . In addition,  $P_n$  and  $C_n$  will respectively represent the traditional path and cycle graphs on  $n$  vertices.

### 3. Basic graphs

We begin by demonstrating our model on some of the simplest possible graphs. Many more complex and interesting structures in graph theory can be constructed

from these basic graphs. For many of these problems the number of rearrangements “with stays”,  $R_s(G)$  is the more interesting problem.

The simplest graph we consider is the path graph on  $n$  vertices. By comparison with the Fibonacci tilings of a  $1 \times n$  board it is easy to see that  $R_s(P_n) = f_n$ . Similarly, we can construct a natural correspondence between Lucas tilings and rearrangements on  $C_n$  that accounts for all rearrangements except the two oriented cycles where each marker moves in the same direction. Thus,  $R_s(C_n) = l_n + 2$ .

Counting the rearrangements on the complete graph of order  $n$  is also a simple counting problem. Considering each rearrangement as a permutation, we see that if each marker must move to a new vertex we have that  $R(K_n)$  is equal to the  $n$ -th derangement number, while if the markers are permitted to stay we have  $R_s(K_n) = n!$ .

Rearrangements on complete bipartite graphs are slightly more complex, yet still yield nice closed form representations.

**Proposition 6.** *The number of rearrangements on  $K_{n,n}$  is equal to  $(n!)^2$ .*

*Proof.* We begin by coloring the vertices of  $K_{n,n}$  black or white according to the bipartition. To construct a rearrangement on  $K_{n,n}$  we note that much like the rectangular classroom problem, we can consider the movements of all of the vertices in each bipartition independently. Without loss of generality, we may order the white vertices. Then, the first white marker may move to any of  $n$  black vertices, while the  $k$ -th white marker can select any of the  $n - k + 1$  remaining black vertices. A similar method can be independently applied to the markers initially placed on black vertices.

Thus, the number of rearrangements of the markers that begin on a particular color is equal to  $\prod_{i=1}^n (n - i + 1) = n!$ . Hence,  $R(K_{n,n}) = (n!)^2$ .  $\square$

In order to simplify the statement of the following result, we define some additional notation. Specifically, let  $(n)_i = n(n - 1)(n - 2) \cdots (n - i + 1)$  represent the standard falling factorial.

**Proposition 7.** *The number of rearrangements with stays on  $K_{m,n}$  is equal to  $\sum_{i=0}^m (m)_i (n)_i$ .*

*Proof.* Without loss of generality we can assume that  $m \leq n$  and color the vertices in the  $m$  partition white and the vertices in the  $n$  partition black. We can count the rearrangements by conditioning on the number of markers that move from a white vertex to a black vertex. Let  $i$  represent the number of markers that move from white to black. Then there are  $\binom{m}{i}$  ways to choose which white markers to move.

For any  $1 \leq k \leq i$  the  $k$ -th moving white marker may select to move to any of  $n - k + 1$  black vertices. This gives us  $(n)_i$  ways to move the  $\binom{m}{i}$  selected white markers.

Graph	Rearrangements	With stays
$P_n$	0, 1, 0, 1, 0, ...	$f_n$
$C_n$	0, 1, 2, 4, 2, 4, ...	$l_n + 2 = f_n + f_{n-2} + 2$
$K_n$	$D(n)$	$n!$
$K_{n,n}$	$(n!)^2$	$\sum_{i=0}^n ((n)_i)^2$
$K_{m,n}$ with $m < n$	0	$\sum_{i=0}^m (m)_i (n)_i$

**Table 1.** Rearrangements on basic graphs.

At this point there are  $i$  empty white vertices and  $i$  black vertices that contain a marker that must be moved. There are  $i!$  ways to construct a legitimate rearrangement from this scenario. Summing over all possible  $i \leq m$  gives

$$\begin{aligned} \sum_{i=0}^m \binom{m}{i} (n)_i i! &= \sum_{i=0}^m \frac{m!}{i!(m-i)!} (n)_i i!, \\ &= \sum_{i=0}^m (m)_i (n)_i. \end{aligned} \quad \square$$

Table 1 summarizes the results of this section, some of which will be referenced later in this paper.

#### 4. Theorems

In this section we present some theoretical results related to our seating rearrangement model. The first theorem generalizes our earlier results on the original rectangular seating rearrangement problem and  $R(K_{n,n})$ .

**Theorem 8.** *Let  $G = (\{U, V\}, E)$  be a bipartite graph. The number of rearrangements on  $G$  is equal to the square of the number of perfect matchings on  $G$ .*

*Proof.* We may construct a bijection between pairs of perfect matchings on  $G$  and cycle covers on  $\vec{G}$ . Without loss of generality, select two perfect matchings of  $G$ ,  $m_1$  and  $m_2$ . For each edge  $(u_1, v_1)$  in  $m_1$ , place a directed edge in the cycle cover from  $u_1$  to  $v_1$ . Similarly, for each edge  $(u_2, v_2)$  in  $m_2$ , place a directed edge in the cycle cover from  $v_2$  to  $u_2$ . Since  $m_1$  and  $m_2$  are perfect matchings, by construction, each vertex in the cycle cover has in-degree and out-degree equal to 1.

Given a cycle cover  $C$  on  $\vec{G}$  construct two perfect matchings on  $G$  by taking the directed edges from vertices in  $U$  to vertices in  $V$  separately from the directed edges from  $V$  to  $U$ . Each of these sets of (undirected) edges corresponds to a perfect matching by the definition of cycle cover and the bijection is complete.

Since there is a one-to-one correspondence between cycle covers on  $\vec{G}$  and rearrangements on  $G$ , the theorem is proved.  $\square$

Our next result considers the case where we are counting the number of rearrangements with stays on a bipartite graph.

**Theorem 9.** *The number of rearrangements on a bipartite graph  $G$ , when the markers on  $G$  are permitted to remain on their vertices, is equal to the number of perfect matchings on  $P_2 \times G$ .*

*Proof.* Observe that  $P_2 \times G$  can be considered as two identical copies of  $G$  where each vertex is connected to its copy by a single edge. To construct a bijection between cycle covers on  $G$  and perfect matchings on  $P_2 \times G$ , associate each self-loop in a cycle cover with an edge between a vertex and its copy in the perfect matching.

Since the graph is bipartite, the remaining cycles in the cycle cover can be decomposed into matching edges from  $U$  to  $V$  and from  $V$  to  $U$  as in the previous theorem.  $\square$

Applying Theorem 9 to the original problem of seating rearrangements gives that the number of rearrangements in a  $m \times n$  classroom, where the students are allowed to remain in place or move, is equal to the number of perfect matchings in  $P_2 \times P_m \times P_n$ . The  $2 \times n$  case is included in the OEIS as A006253 [OEIS 2012]. These matchings are equivalent to tiling a  $2 \times m \times n$  rectangular prism with  $1 \times 1 \times 2$  tiles. This is a well-known problem that is contained in books on combinatorics, for example [Graham et al. 1994].

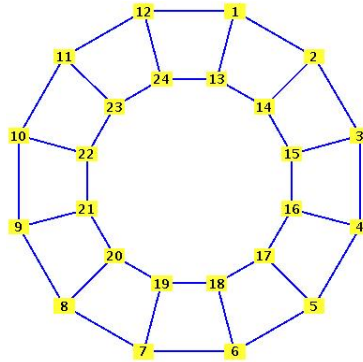
A more direct proof of this equivalence between rectangular seating rearrangements with stays and three-dimensional tilings can be given by associating each possible student move type—up/down, left/right, or remain in place—with a particular tile orientation in space. Then, a tiling can be directly constructed from a given seating rearrangement in a one-to-one fashion.

## 5. Counting examples

We conclude by presenting some examples of the types of counting problems that may be generated with this model. Especially noteworthy are the number of different techniques that may be used to solve these problems.

**5.1. Prism graphs.** The prism graph of order  $n$ , denoted prism  $n$ , is isomorphic to  $C_n \times P_2$ . Rearrangements on prism  $n$  can be considered as  $2 \times n$  classroom seating rearrangements on a cylinder.

**Example 10.** *The number of rearrangements on prism  $n$  is equal to  $(l_n + 2)^2$  when  $n$  is even.*



**Figure 4.** The prism graph of order 12.

*Proof.* Let  $n$  be an even natural number. Then it is easy to see that prism  $n$  is a bipartite graph, since each  $C_n$  is bipartite, and any cycle that includes edges in both  $C_n$  must also be of even length. Since the graph is bipartite, by Theorem 8, the number of rearrangements is equal to the square of the number of perfect matchings. Furthermore, by Theorem 9, the number of perfect matchings on  $C_n \times P_2$  is equal to the number of rearrangements with stays on  $C_n$ , which we showed in Section 3 was equal to  $l_n + 2$ . Squaring this quantity gives the result.  $\square$

**Example 11.** *The number of rearrangements on prism  $n$  is equal to  $l_{2n} + 2$  when  $n$  is odd.*

*Proof.* Let  $n$  be an odd natural number. In this case prism  $n$  is not bipartite, so we must make a different argument. First note that we can divide the rearrangements into two classes by whether a marker moves between the two  $C_n$  in the rearrangement. There are exactly four rearrangements for each  $n$  where no markers move between the two  $C_n$ , as these correspond to simple cycles where each marker on a  $C_n$  moves exactly one square in one direction.

The remaining rearrangements can be placed into a bijection with two independently selected Lucas tilings of order  $n$  where a square in a Lucas tiling represents a move between the  $C_n$ . Note that since  $n$  is odd, any Lucas tiling of order  $n$  must contain at least one square so we are not counting the rearrangements in the first class twice.

Combining these two cases, we have  $R(\text{prism } n) = l_n^2 + 4$ . Using a well-known Lucas identity we can simplify this expression as:

$$l_n^2 + 4 = (l_n^2 + 2) + 2 = l_{2n} + 2. \quad \square$$

Computing the number of rearrangements with stays on a prism graph is a much more difficult problem. Considering all of the possible ways to rearrange

$n$	3	4	5	6	7	8
No stays	20	81	125	400	845	2401
With stays	82	272	890	3108	11042	39952
$n$	9	10	11	12	13	14
No stays	5780	15625	39605	104976	271445	714025
With stays	146026	537636	1988722	7379216	27436250	102144036

**Table 2.** Rearrangements on prism graphs.

an arbitrary pair of adjacent markers each in a separate  $C_n$  gives a system of 11 homogeneous, linear recurrence relations. This system is fully derived and demonstrated in Appendix A. This system can then be solved, using the successor operator method due to DeTemple and Webb [2014], to give the following solution:

$$a_n = 10a_{n-1} - 36a_{n-2} + 50a_{n-3} + 11a_{n-4} - 108a_{n-5} + 96a_{n-6} \\ + 20a_{n-7} - 75a_{n-8} + 34a_{n-9} + 4a_{n-10} - 6a_{n-11} + a_{n-12},$$

with initial conditions given in Table 2.

Using these initial conditions we were further able to construct a generalized power sum by solving a linear equation in the eigenvalues of the recurrence to determine the coefficients:

$$R_s(\text{prism } n) = 6 + 4(-1)^n + (2 + \sqrt{3})^n + (2 - \sqrt{3})^n + 2(1 + \sqrt{2})^n + 2(1 - \sqrt{2})^n.$$

Since the repeated eigenvalues have coefficients of 0 in the generalized power sum, our sequence must also satisfy a recurrence of order 6. By computing the implied characteristic polynomial, we get the following minimal recurrence for this sequence:

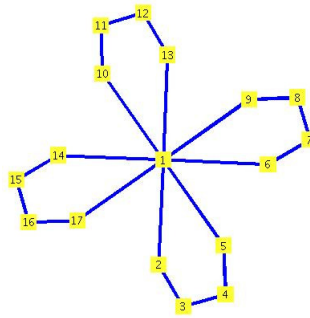
$$a_n = 6a_{n-1} - 7a_{n-2} - 8a_{n-3} + 9a_{n-4} + 2a_{n-5} - a_{n-6}.$$

**5.2. Dutch windmills.** A Dutch windmill,  $Dw_n^m$ , consists of  $m$  copies of an  $n$  cycle all joined at a single vertex. For example, the friendship graphs are  $F_k = Dw_3^k$ . Counting the rearrangements on Dutch windmills highlights some of the Fibonacci relations of these counting problems.

**Example 12.** *The number of rearrangements on  $Dw_n^m$  is 0 when  $n$  is even and  $2m$  when  $n$  is odd.*

*Proof.* We may condition on the movement of the marker initially positioned on the center vertex. The center vertex is adjacent to  $2m$  other vertices, and every rearrangement on  $Dw_n^m$  must consist of a single  $n$ -cycle containing the center vertex





**Figure 5.** A Dutch windmill  $Dw_5^4$ .

and  $(m - 1)(n - 1)/2$  two-cycles pairing up the remaining vertices as there are no other cycles remaining in the graph.

When  $n$  is even, removing the center vertex from all but one of the  $n$ -cycles leaves an odd number of vertices, which cannot be satisfactorily paired together. Thus, there can be no legitimate rearrangements when  $n$  is even.

In the case where  $n$  is odd, the movement of the center marker onto one of its  $2m$  neighbors completely determines the rearrangement. □

**Example 13.** *The number of rearrangements with stays permitted on  $Dw_n^m$  is  $(f_{n-1})^m + 2m(f_{n-2} + 1)(f_{n-1})^{m-1}$ .*

*Proof.* We may again condition on the behavior of the center marker. There are two cases: either the center marker moves to an adjacent vertex or it remains in place.

When the center marker does not move, the remaining markers form  $m$  copies of  $P_{n-1}$ , which may each be rearranged independently in  $f_{n-1}$  ways.

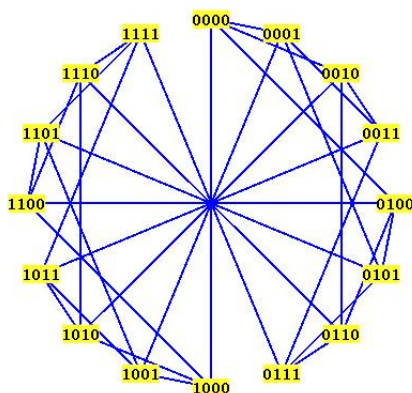
When the center marker moves onto one of the  $2m$  adjacent vertices, it either lies on a two-cycle, in which case there are  $f_{n-2}$  ways for the other vertices on that cycle to rearrange themselves, or it lies on the entire  $n$ -cycle. The  $m - 1$  remaining  $n$ -cycles that were not selected are again each reduced to  $P_{n-1}$ , contributing  $(f_{n-1})^{m-1}$  to the rearrangement total.

Combining these two cases gives the desired result:

$$R_S(Dw_n^m) = (f_{n-1})^m + 2m(f_{n-2} + 1)(f_{n-1})^{m-1}. \quad \square$$

**5.3. Hypercubes.** Hypercubes are a commonly studied mathematical object, and enumerating the perfect matchings on an arbitrarily large hypercube is an open problem in combinatorics [Lundow 1996]. Rearrangements, both with and without stays, have interesting connections to this problem.

The hypercube of order  $n$  can be constructed as a graph whose vertices are labeled with the  $2^n$  binary strings of length  $n$ , with an edge between two vertices



**Figure 6.** The hypercube of order 4.

	1	2	3	4	5
$R(H_n)$	1	4	81	73984	347138964225
$R_s(H_n)$	2	9	272	589185	16332454526976

**Table 3.** Hypercube rearrangements.

when the respective labels differ in only one location. More importantly for our purposes, if  $H_n$  represents the hypercube of order  $n$ , then  $H_n \cong H_{n-1} \times P_2$ .

Thus, the relations below follow directly from Theorem 8 and Theorem 9.

**Corollary 14.** *The number of rearrangements on  $H_n$  is equal to the square of the number of perfect matchings on  $H_n$ .*

**Corollary 15.** *The number of rearrangements with stays on  $H_n$  is equal to the number of perfect matchings on  $H_{n+1}$ .*

### Appendix A. Computing $R_s(\text{prism } n)$

In this appendix, we give the full derivation of the generalized power sum for  $R_s(\text{prism } n)$ . Recall that prism  $n$  is isomorphic to  $C_n \times P_2$  and may be considered a discrete  $2 \times n$  cylinder. Thus, this problem is equivalent to the original seating rearrangement problem in a cylindrical classroom. Our goal is to construct a system of linear recurrences representing the ways that the (arbitrarily chosen) first column of desks can be filled.

We begin by letting  $a_n$  represent the number of rearrangements on prism  $n$ . Figure 7 shows all of the possible endings that we need to account for in our system. The dots in the figure represent students that have not moved, while the crosses represent students that have already moved. Note that the endings are representations

of classes of endings, up to symmetry. Thus, for example, an ending counts as a  $c_n$  regardless of whether the completed desk is in the top or bottom row, since the number of rearrangements is the same. To see how the system is constructed consider the possible movements of the students in the first column of a  $b_n$ :

- The two students may elect to either remain in their seats or swap seats with each other; either of these choices leaves a  $b_{n-1}$ .
- Both students may swap seats with the next student in their row, leaving a  $b_{n-2}$ .
- One of the students may remain in his seat, while the other swaps with his horizontal neighbor. This can happen in two ways, so we have  $2c_{n-1}$ .
- One of the students may move vertically, while the other moves horizontally. Again this can happen in two ways, and our sum gains a term of  $2d_{n-1}$ .

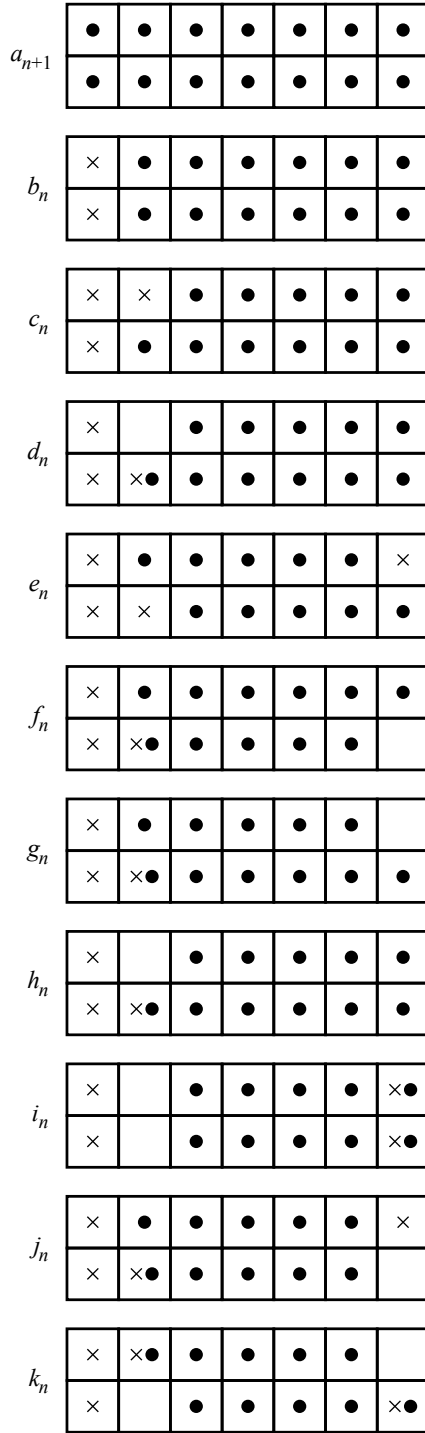
Similarly, consider the possibilities for a classroom ending set-up as  $g_n$ . As shown in Figure 7, we will assume that the desk with two students is in the upper left while the empty desk is in the lower right. However, this analysis extends to any rotation or reflection of  $g_n$ .

- The student yet to move in the upper left may move vertically forcing the student in the bottom left to move horizontally. This leaves a  $f_{n-1}$ .
- The student yet to move in the upper left may move horizontally while the student below remains in place. The remaining situation is a  $g_{n-1}$ .
- The student yet to move in the upper left may move horizontally while the student below swaps places horizontally, which forces a  $g_{n-2}$ .

Extending this reasoning to all of the endings under consideration leads to the following system of recurrences:

$$\begin{aligned}
 a_n &= 2b_{n-1} + 2b_{n-2} + 4c_{n-1} + 2e_{n-1} \\
 &\quad + 4f_{n-1} + 4g_{n-1} + 4h_{n-1} + 2i_{n-1} + 2j_{n-1} + 2k_{n-1}, \\
 b_n &= 2b_{n-1} + b_{n-2} + 2c_{n-1} + 2d_{n-1}, & c_n &= b_{n-1} + c_{n-1}, \\
 d_n &= b_{n-1} + d_{n-1}, & e_n &= c_{n-1} + e_{n-1}, \\
 f_n &= f_{n-1} + f_{n-2} + g_{n-1}, & g_n &= f_{n-1} + g_{n-1} + g_{n-2}, \\
 h_n &= b_{n-1} + h_{n-1}, & i_n &= i_{n-1}, \\
 j_n &= f_{n-1}, & k_n &= k_{n-1} + h_{n-1}.
 \end{aligned}$$

Applying the successor operator,  $E$ , to this system gives us the following symbolic matrix whose determinant is the characteristic polynomial of the recurrence relation we are seeking.



**Figure 7.** Prism endings.

$M =$

$$\begin{bmatrix} E^2 & -2E-2 & -4E & 0 & -2E & -4E & -4E & -4E & -2E & -2E & -2E \\ 0 & E^2-2E-1 & -2E & -2E & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & E-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & E-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & E-1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & E^2-E-1 & -E & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -E & E^2-E-1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & E-1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & E-1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & E & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & E-1 \end{bmatrix}$$

Note that  $M$  is defined to satisfy the following equation as the successor operator acts on each sequence in turn:

$$M [a_n \ b_n \ c_n \ d_n \ e_n \ f_n \ g_n \ h_n \ i_n \ j_n \ k_n]^T = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]^T$$

We can now calculate the determinant of  $M$  and construct our recurrence relation,

$$\det M = E^{15} - 10E^{14} + 36E^{13} - 50E^{12} - 11E^{11} + 108E^{10} - 96E^9 - 20E^8 + 75E^7 - 34E^6 - 4E^5 + 6E^4 - E^3.$$

The coefficients of this characteristic polynomial give us our first recurrence relation (Section 5), while the roots of the polynomial are the eigenvalues of our recurrence. After removing the zeros, these eigenvalues and their multiplicities are  $\{1^6, -1^2, 1 + \sqrt{2}, 1 - \sqrt{2}, 2 + \sqrt{3}, 2 - \sqrt{3}\}$ . Thus, our characteristic polynomial factors to:

$$E^3(E - 1)^6(E + 1)^2(E^2 - 4E + 1)(E^2 - 2E - 1).$$

To find the generalized power sum, we solve the linear system  $Ax = b$ , where  $A$  represents the eigenvalues matrix (with elements multiplied by powers of  $n$  where necessary to preserve linear independence),  $x$  represents the coefficients vector, and  $b$  the initial conditions as shown in Table 2. The coefficients obtained as a solution to this system give the generalized power sum described previously in Section 5.

Taking a product of only the factors corresponding to the eigenvalues in the generalized power sum gives us the following characteristic polynomial of degree 6:

$$(E^2 - 4E + 1)(E^2 - 2E - 1)(E - 1)(E + 1) = E^6 - 6E^5 + 7E^4 + 8E^3 - 9E^2 - 2E + 1.$$

Since this polynomial also annihilates our sequence, its corresponding recurrence relation must also be satisfied by our sequence. This gives the second recurrence relation in Section 5. By exhaustively examining the factors of this polynomial we find that it is the polynomial of minimal degree that represents our sequence.

### Acknowledgements

I would like to express my gratitude towards Dr. William Webb for his assistance and insight. This work was supported by a grant from the Washington State University College of Sciences.

### References

- [Aaronson 2011] S. Aaronson, “A linear-optical proof that the permanent is #P-hard”, *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **467**:2136 (2011), 3393–3405. MR 2853286 Zbl 06062301
- [Beineke and Harary 1966] L. W. Beineke and F. Harary, “Binary matrices with equal determinant and permanent”, *Studia Sci. Math. Hungar* **1** (1966), 179–183. MR 34 #7397 Zbl 0145.01505
- [Benjamin and Quinn 2003] A. T. Benjamin and J. J. Quinn, *Proofs that really count: The art of combinatorial proof*, The Dolciani Mathematical Expositions **27**, Mathematical Association of America, Washington, DC, 2003. MR 2004f:05001 Zbl 1044.11001
- [Chartrand et al. 2011] G. Chartrand, L. Lesniak, and P. Zhang, *Graphs & digraphs*, 5th ed., CRC Press, Boca Raton, 2011. MR 2012c:05001 Zbl 1211.05001
- [DeTemple and Webb 2014] D. DeTemple and W. Webb, *Combinatorial reasoning: An introduction to the art of counting*, Wiley, Hoboken, NJ, 2014.
- [Graham et al. 1994] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete mathematics: A foundation for computer science*, 2nd ed., Addison-Wesley, Reading, MA, 1994. MR 97d:68003 Zbl 0836.00001
- [Harary 1967] F. Harary, *Graph theory and theoretical physics*, Academic Press, New York, 1967. MR 37 #6208 Zbl 0202.55602
- [Harary 1969] F. Harary, “Determinants, permanents and bipartite graphs”, *Math. Mag.* **42**:3 (1969), 146–148. MR 39 #4035 Zbl 0273.15006
- [Honsberger 1997] R. Honsberger, *In Pólya’s footsteps: Miscellaneous problems and essays*, The Dolciani Mathematical Expositions **19**, Mathematical Association of America, Washington, DC, 1997. MR 98d:00002 Zbl 0893.00005
- [Kasteleyn 1961] P. Kasteleyn, “The statistics of dimers on a lattice, I: The number of dimer arrangements on a quadratic lattice”, *Physica* **27**:12 (1961), 1209–1225. Zbl 1244.82014
- [Kennedy and Cooper 1993] R. Kennedy and C. Cooper, “Variations on a  $5 \times 5$  seating rearrangement problem”, *Mathematics in College* **1993** (1993), 59–67.
- [Kuperberg 1998] G. Kuperberg, “An exploration of the permanent-determinant method”, *Electron. J. Combin.* **5** (1998), R46, 1–34. MR 99j:05141 Zbl 0906.05055
- [van Lint and Wilson 2001] J. H. van Lint and R. M. Wilson, *A course in combinatorics*, 2nd ed., Cambridge University Press, 2001. MR 2002i:05001 Zbl 0980.05001
- [Little 1975] C. H. C. Little, “A characterization of convertible  $(0,1)$ -matrices”, *J. Combinatorial Theory Ser. B* **18**:3 (1975), 187–208. MR 54 #12542 Zbl 0281.05013
- [Loehr 2011] N. A. Loehr, *Bijjective combinatorics*, CRC Press, Boca Raton, 2011. MR 2012d:05002 Zbl 1234.05001
- [Lundow 1996] P. Lundow, “Computation of matching polynomials and the number of 1-factors in polygraphs”, research Reports 12, Umeå University, 1996, <http://www.theophys.kth.se/~phl/Text/1factors.pdf>.

- [Marcus and Minc 1965] M. Marcus and H. Minc, “Permanents”, *Amer. Math. Monthly* **72** (1965), 577–591. MR 31 #1266 Zbl 0166.29904
- [OEIS 2012] OEIS, “The On–Line Encyclopedia of Integer Sequences”, 2012, <http://oeis.org>.
- [Otake et al. 1996] T. Otake, R. Kennedy, and C. Cooper, “On a seating rearrangement problem”, *Mathematics and Informatics Quarterly* **52** (1996), 63–71.
- [Pólya 1913] G. Pólya, “Aufgabe 424”, *Arch. Math. Phys.* **20**:3 (1913), 271.
- [Robertson et al. 1999] N. Robertson, P. D. Seymour, and R. Thomas, “Permanents, Pfaffian orientations, and even directed circuits”, *Ann. of Math. (2)* **150**:3 (1999), 929–975. MR 2001b:15013 Zbl 0947.05066
- [Shilov 1977] G. E. Shilov, *Linear algebra*, Dover, New York, 1977. MR 57 #6043 Zbl 0218.15003
- [Temperley and Fisher 1961] H. N. V. Temperley and M. E. Fisher, “Dimer problem in statistical mechanics—an exact result”, *Philos. Mag. (8)* **6**:68 (1961), 1061–1063. MR 24 #B2436 Zbl 0126.25102
- [Valiant 1979] L. G. Valiant, “The complexity of computing the permanent”, *Theoret. Comput. Sci.* **8**:2 (1979), 189–201. MR 80f:68054 Zbl 0415.68008
- [Vazirani and Yannakakis 1988] V. V. Vazirani and M. Yannakakis, “Pfaffian orientations, 0/1 permanents, and even cycles in directed graphs”, pp. 667–681 in *Automata, languages and programming* (Tampere, 1988), edited by T. Lepistö and A. Salomaa, Lecture Notes in Comput. Sci. **317**, Springer, Berlin, 1988. MR 90k:68078 Zbl 0648.68060

Received: 2013-11-04    Revised: 2014-01-03    Accepted: 2014-01-24

ddeford@math.dartmouth.edu    *Department of Mathematics, Dartmouth College,  
27 North Main Street, Hanover, NH 03755, United States*





# Fibonacci Nim and a full characterization of winning moves

Cody Allen and Vadim Ponomarenko

(Communicated by Scott T. Chapman)

In this paper we will fully characterize all types of winning moves in the “take-away” game of Fibonacci Nim. We prove the known winning algorithm as a corollary of the general winning algorithm and then show that no other winning algorithms exist. As a by-product of our investigation of the game, we will develop useful properties of Fibonacci numbers. We conclude with an exploration of the probability that unskilled player may beat a skilled player and show that as the number of tokens increase, this probability goes to zero exponentially.

## 1. Introduction

We begin with a brief introduction to the idea of “take-away” games. Schwenk [1970] defined *take-away* games to be a two-person game in which the players alternately diminish an original stock of tokens subject to various restrictions, with the player who removes the last token being the winner.

In the generalized *take-away* game,  $\tau(k) = \eta(k-1) - \eta(k)$  where  $\eta(k)$  is the number of tokens remaining after the  $k$ -th turn so that  $\tau(k)$  is the number of tokens removed on the  $k$ -th turn. Additionally, for all  $k \in \mathbb{N}$ ,  $k \neq 1$ , we have  $\tau(k) \leq m_k$ , where  $m_k$  is some function of  $\tau(k-1)$ . Specifically in Fibonacci Nim, we have  $m_k = 2\tau(k)$  for  $k > 1$ . We will immediately move away from this notation and develop additional notation as it is required. We provide a simple example to familiarize the reader with the game.

**Example 1.** Let  $n = 10$ . Player one may remove 1 to 9 tokens. Suppose player one removes 3 tokens. Then, player two may now remove 1 through  $2(3) = 6$  tokens. Play continues until one of the players removes the last token.

We will rely heavily on results from [Lekkerkerker 1952], specifically the *Zeckendorff representation* of natural numbers as a sum of Fibonacci numbers.

---

*MSC2010:* primary 91A46; secondary 11A63.

*Keywords:* Fibonacci Nim, take away games, dynamic Nim, combinatorial games, Fibonacci.

The Fibonacci numbers are the positive integers generated by the recursion  $F_k = F_{k-1} + F_{k-2}$ , where  $F_1 = 1 = F_2$  and  $k \in \mathbb{N}$ . Let  $F = \{F_2, F_3, \dots, F_k, \dots\} = \{1, 2, 3, 5, \dots\}$ . This is the subset of Fibonacci numbers we will reference throughout this paper. We now present the Zeckendorf representation theorem (ZRT) without proof. A proof of this theorem may be found in [Hoggatt et al. 1973].

**Theorem 2** (Zeckendorff representation theorem). *Let  $n \in \mathbb{N}$ . For  $i, r \in \mathbb{N}$  we have  $n = F_{i_r} + F_{i_{r-1}} + \dots + F_{i_1}$ , where  $i_r - (r - 1) > i_{r-1} - (r - 2) > \dots > i_2 - 1 > i_1 \geq 2$ . Further, this representation is unique.*

In other words, every positive integer can be written uniquely as a sum of non-consecutive Fibonacci numbers. Clearly, in the notation of the theorem,  $F_{i_r} > F_{i_{r-1}} > \dots > F_{i_1}$ . We will refer to the Zeckendorff Representation theorem frequently, so we abbreviate it by ZRT.

**Example 3.**  $12 = (1)F_6 + (0)F_5 + (1)F_4 + (0)F_3 + (1)F_2 = 8 + 3 + 1$ .

**Corollary 4.** *If  $F_{k+1} > n \geq F_k$ , then  $F_k$  is the largest number in the Zeckendorff representation of  $n$ .*

*Proof.* If  $F_{k+1} > n \geq F_k$  then by Zeckendorff’s theorem we can write  $(n - F_k) = F_d + \dots + F_{i_1}$ . We claim  $k - 1 > d$ . Suppose not; then  $d \geq (k - 1)$ , thus

$$n = F_k + F_d + \dots + F_{i_1} \geq F_k + F_d \geq F_k + F_{k-1} = F_{k+1}.$$

However,  $F_{k+1} > n \geq F_{k+1}$  is a contradiction. Thus,  $k - 1 > d$  so that

$$n = F_k + F_d + \dots + F_{i_1}$$

is a valid, and thus the only, representation of  $n$  by the ZRT. □

The corollary above shows that for any  $n \in \mathbb{N}$  where  $F_{k+1} > n \geq F_k$ , the Zeckendorff representation of  $n$  must contain  $F_k$ . Therefore, we iteratively may take the maximal Fibonacci number less than  $n$ , say  $F_k$ , subtract it from  $n$  which yields  $n - F_k = n' = F_{i_{r'}} + F_{i_{r'-1}'} + \dots + F_{i_1}'$  and repeat this process to find each Fibonacci number in the representation of the original number,  $n$ .

**Definition 5.** Let  $n = F_{i_r} + F_{i_{r-1}} + \dots + F_{i_1}$  where  $r, i, n \in \mathbb{N}$ . We define  $T(n) = F_{i_1}$ . That is,  $T(n)$  is the smallest number in the Zeckendorff representation.

**Definition 6.** Let  $n = F_{i_r} + F_{i_{r-1}} + \dots + F_{i_1}$  where  $r, i, n, j \in \mathbb{N}$ . We now define the *length  $j$  tail* to be the specific sum of  $j$  consecutive\* Fibonacci numbers in the Zeckendorff representation of  $n$  beginning with the smallest number,  $F_{i_1}$ . We set  $T_1(n) = T(n)$  for consistency. Then,  $T_j(n) = T(n) + T_{j-1}(n - T(n))$ .

The “consecutive\*” in Definition 6 refers to the subscripts  $i_j, i_{j+1}$  for some  $j \in \mathbb{N}$ . By the above definitions, we see that the *length  $j$  tail* of  $n$  is  $T_j(n) = F_{i_j} + F_{i_{j-1}} + \dots + F_{i_1}$  where  $r \geq j \geq 1$ .

**Example 7.** Consider

$$\begin{aligned} 33 &= F_8 + F_6 + F_4 + F_2 = 21 + 8 + 3 + 1, \\ 12 &= F_6 + F_4 + F_2 = 8 + 3 + 1. \end{aligned}$$

Then, the *length 3 tails* are

$$\begin{aligned} T_3(33) &= F_6 + F_4 + F_2 = 8 + 3 + 1, \\ T_3(12) &= F_6 + F_4 + F_2 = 8 + 3 + 1. \end{aligned}$$

Hence, 33 and 12 have the same length 3 tail.

**Remark 8.** By the definition of a length  $j$  tail, if  $T_j(n) = T_j(m)$ , then for any  $j \geq s \geq 1$ , we have  $T_s(n) = T_s(m)$ .

Let  $n = F_{i_r} + F_{i_{r-1}} + \dots + F_{i_1}$  be the Zeckendorff representation where

$$F_{i_r} > F_{i_{r-1}} > \dots > F_{i_1}.$$

Suppose there are  $n$  tokens in the pile during the current turn. The known winning algorithm for Fibonacci Nim has the current player take the length 1 tail of  $n$ . That is, the player removes  $T(n) = F_{i_1}$  tokens. We will prove that this is a winning algorithm in the next section.

In what follows, we will extend the known winning algorithm to include tails that satisfy certain criteria for some given  $n$ . We then will prove that this is a complete collection of winning moves and that no others exist. We end this paper by introducing a *losing position strategy* and then derive an upperbound on the probability that an unskilled player may beat a skilled player.

## 2. Fibonacci Nim strategy

We begin this section by discussing how to win Fibonacci Nim. In the remainder of this paper, we use  $n = F_{i_r} + F_{i_{r-1}} + \dots + F_{i_1}$  with  $n, r, i \in \mathbb{N}$  as the Zeckendorff representation for some  $n$ .

Assume there are  $n$  tokens in a given turn which the player whose turn it is may remove from. Let  $2p$  denote the maximum number of tokens this player may remove from the  $n$  tokens. We can denote this position by  $(n, 2p)$ . Note, this implies that the previous player removed precisely  $p$  tokens.

**Definition 9.** A *losing position* is such that given the position  $(n, 2p)$ ,  $T(n) > 2p$ . A *winning position* is any nonlosing position. A *winning move* is such that it results in the next position being a losing position. A *losing move* is any nonwinning move.

We see by Definition 9 that we always want to leave our opponent in a losing position where  $T(n) > 2p$ . That is, a position where our opponent cannot remove any length  $j$  tail,  $T_j(n)$ . As an immediate consequence, if our opponent cannot remove

a tail of  $n$ , certainly he cannot remove all of  $n$  to win since  $n \geq T_j(n) \geq T(n) > 2p$ . Therefore, if we can successively give our opponent a losing position, we can ensure we win.

**Lemma 10.** *For every  $i \in \mathbb{N}$  where  $i \geq 3$ ,  $2F_{i-1} \geq F_i$  and  $F_{i+1} > 2F_{i-1}$ .*

*Proof.* We have  $2(F_2) = 2(1) = 2 = F_3$  and  $F_4 = 3 > 2 = 2(1) = 2(F_2)$ . Assume that  $2F_{i-1} \geq F_i$  and  $F_{i+1} > 2F_{i-1}$ . We have  $2(F_i) = 2(F_{i-1} + F_{i-2}) \geq 2F_{i-1} + F_{i-2} = F_i + F_{i-1} = F_{i+1}$  since for each for  $j \in \mathbb{N}$ ,  $F_j \geq 1$ . Similarly,

$$\begin{aligned} F_{i+2} &= F_{i+1} + F_i = (F_i + F_{i-1}) + (F_{i-1} + F_{i-2}) \\ &> F_i + (F_{i-1} + F_{i-2}) = 2F_i. \end{aligned} \quad \square$$

Lemma 11 below implies that if on a given turn our opponent has a losing position to play from, regardless of how he plays, our next play will be from a winning position.

**Lemma 11.** *Let  $n \in \mathbb{N}$ . For any  $p$  with  $T(n) > p$ ,  $(n - p, 2p)$  is a winning position.*

*Proof.* Let  $n \in \mathbb{N}$ . Assume  $T(n) > p$ . We have,  $n - p = F_{i_r} + \dots + F_{i_1} - p$ . Define  $m = T(n) - p = F_{i'_r} + \dots + F_{i'_1}$ . Suppose  $(n - p, 2p)$  is a losing position. Then,  $T(n - p) > 2p$  and by Lemma 10,  $2F_{i'_1-1} \geq F_{i'_1} > 2p$ . Hence, the Zeckendorff representation of  $p$  does not include  $F_{i'_1-1}$ , thus  $p = F_{i''_r} + \dots + F_{i''_1}$  where  $F_{i'_1-1} > F_{i''_1}$ . But then,  $n = F_{i_r} + \dots + F_{i_2} + F_{i_r} + \dots + F_{i'_1} + F_{i''_r} + \dots + F_{i''_1}$  is a valid Zeckendorff representation of  $n$ . This is a contradiction since Zeckendorff representations are unique. Hence, we must have  $2p \geq T(n - p)$ . Since  $F_{i_2} > T(n) > T(n) - p$ , then,  $n - p = F_{i_r} + \dots + F_{i_2} + m$  is a valid Zeckendorff representation of  $n - p$  and hence the only representation. Thus, the next position,  $(n - p, 2p)$  has  $2p \geq T(n - p)$  so that  $(n - p, 2p)$  is a winning position.  $\square$

Lemma 12 below paired with Lemma 11 proves the known winning strategy. That is, if we take the length 1 tail of  $n$ ,  $T(n)$ , the next position is a losing position. Successively implementing this lemma results in winning the game in a finite number of moves.

**Lemma 12.** *Let  $n \in \mathbb{N}$ . Set  $p = T(n)$ . Then  $(n - p, 2p)$  is a losing position.*

*Proof.* Let  $n \in \mathbb{N}$ . Set  $p = T(n)$ . Suppose for some  $k \in \mathbb{N}$ ,  $F_k = T(n) = F_{i_1}$ . By Theorem 2,  $F_{i_2} \geq F_{k+2}$ . Then, by Lemma 10,  $F_{i_2} \geq F_{k+2} > 2F_k = 2p$ . By uniqueness of the ZRT,  $n - p = F_{i_r} + \dots + F_{i_2}$  and  $(n - p, 2p)$  has  $T(n - p) = F_{i_2} > 2p$ . Hence,  $(n - p, 2p)$  is a losing position.  $\square$

For now, we state that not every tail may always be taken from  $n$  to produce a losing position. In the following subsections, we will prove this rigorously and derive results which show exactly which tails may be removed to put our opponent in a losing position. Theorem 13 is this section's main result. Namely, it proves that removing length  $j$  tails of  $n$  are the *only* winning moves for  $n \in \mathbb{N}$ .

**Theorem 13** (Fundamental theorem of Fibonacci Nim). *Let  $n \in \mathbb{N}$ . Then, for any  $p \notin \{T_{r-1}(n), T_{r-2}(n), \dots, T(n)\}$ ,  $(n - p, 2p)$  is a winning position.*

*Proof.* Let  $n \in \mathbb{N}$  and suppose our opponent has removed  $p$  tokens. Then the current position is  $(n - p, 2p)$ . Assume  $T(n - p) > 2p$ , that is,  $(n - p, 2p)$  is a losing position. If  $p = T_j(n)$  for some  $r > j \geq 1$ , then  $p \in \{T_{r-1}(n), T_{r-2}(n), \dots, T(n)\}$ . This leaves two cases to examine: (1)  $p$  is a sum of terms  $F_{i_t}$  where  $r \geq t \geq 1$  and  $p \neq T_j(n)$  for some  $r > j \geq 1$  or (2)  $p \neq T_j(n)$  for some  $r > j \geq 1$  and  $p$  is not of the form given in case (1).

*Case 1:* Our opponent removes  $p = a_r F_{i_r} + a_{r-1} F_{i_{r-1}} + \dots + a_1 F_{i_1}$  where each  $a_j \in \{0, 1\}$  for  $j \in [1, i_r]$  and there exists at least one pair  $(a_j, a_{j+1})$  such that  $a_j = 0$  and  $a_{j+1} = 1$  in the representation of  $p$ . Then,  $p \neq T_j(n)$  for some  $r > j \geq 1$ . Without loss of generality, let  $(a_j, a_{j+1})$  be the minimal pair such that  $a_j = 0$  and  $a_{j+1} = 1$  in the representation of  $p$ . Define

$$n' = (F_{i_r} - a_r F_{i_r}) + \dots + (F_{i_{j+1}} - a_{j+1} F_{i_{j+1}}) + (F_{i_{j-1}} - a_{j-1} F_{i_{j-1}}) + \dots + (F_{i_1} - a_1 F_{i_1}).$$

Then,

$$n - p = F_{i_r} + F_{i_{r-1}} + \dots + F_{i_1} - (a_r F_{i_r} + a_{r-1} F_{i_{r-1}} + \dots + a_1 F_{i_1}) = n' + F_{i_j}$$

which is a valid Zeckendorff representation and hence the only representation of  $n - p$ . Since  $(a_j, a_{j+1})$  is minimal,  $T(n - p) = F_{i_j}$ . We have  $2p > F_{i_{j+1}} > T(n - p)$ , thus  $(n - p, 2p)$  is a winning position and we have reached a contradiction.

*Case 2:* Our opponent removes  $p$  tokens such that  $p \neq a_r F_{i_r} + a_{r-1} F_{i_{r-1}} + \dots + a_1 F_{i_1}$  where each  $a_j \in \{0, 1\}$ . Since  $(n - p, 2p)$  is a losing position, by Lemma 11 we must have  $p > T(n)$ . Without loss, let  $T_j(n)$  for  $r > j \geq 1$  be the minimal tail such that  $p > T_j(n)$ . By assumption,  $p \neq T_j(n)$ . We have  $F_{i_{j+1}} + T_j(n) > p > T_j(n)$  so that  $F_{i_{j+1}} > p - T_j(n) > 0$ . Define  $\delta p = p - T_j(n)$  so that  $p = T_j(n) + \delta p$ . Let  $m = n - T_j(n)$ . Then,  $n - p = m + T_j(n) - (T_j(n) + \delta p) = m - \delta p$ . Since  $T(m) > \delta p$ , by Lemma 11 and the uniqueness of Zeckendorff representations,  $(m - \delta p, 2\delta p)$  is a winning position. It follows that  $2p > 2\delta p \geq T(m - \delta p) = T(n - p)$ . Therefore,  $(n - p, 2p)$  is a winning position and we have reached a contradiction.

Hence, removing some  $p \neq T_j(n)$  for some  $r > j \geq 1$  results in a winning position. Since there is only one other possible move, removing some tail  $F_j(n)$ , it follows that if  $(n - p, 2p)$  is a losing move, then  $p = T_j(n)$  for some  $r > j \geq 1$ .  $\square$

**Remark 14.** By Definition 9 and Theorem 13, removing  $T_j(n)$  tokens where  $r > j \geq 1$  will force an immediate losing position to our opponent when  $F_{i_{j+1}} > 2T_j(n)$ .

In section (2) we have shown that the only possible winning moves in Fibonacci Nim are those that are partial consecutive\* sums or, tails of the Zeckendorff representation of the number of tokens in that turn. In the next section, we determine

which tails force losing positions and how to identify these tails based solely on the Zeckendorff representation for a given  $n$ .

### 3. Winning tails

In this Section, we will show how to take the result from Remark 14: *removing  $T_j(n)$  tokens where  $r > j \geq 1$  will force an immediate losing position to our opponent when  $F_{i_{j+1}} > 2T_j(n)$*  and identify which tails satisfy this condition. Existence of winning moves was proved for *Dynamic One-Pile Nim* in a paper by Holshouser, Reiter and Rudzinski [2003]; Fibonacci Nim is classified as a dynamic one-pile Nim game in their paper. Below, we validate the existence of these moves as well as carefully show exactly how to find these winning moves. In addition, we have included a table at the end of this paper to present these results for the first 90 positive integers.

We are concerned with which tails can be taken and which cannot. That is, if  $n = F_{i_r} + F_{i_{r-1}} + \dots + F_{i_1}$ , when is  $F_{i_{j+1}} > 2T_j(n)$  for  $r > j \geq 1$ ? We accomplish this by looking at an arbitrary tail  $T_j(n)$  of  $n$ . We classify exactly when taking  $T_j(n)$  results in leaving a losing position to our opponent.

We begin by setting  $a_{j+1} = i_{j+1} - i_j$  and  $a_j = i_j - i_{j-1}$ . Then,  $a_{j+1}$  and  $a_j$  are the differences in the subscripts of consecutive\* Fibonacci numbers in a Zeckendorff representation of  $n$ . In this section we will show that for any  $F_{i_j}$ , by considering the “gaps” around it, where the gaps are the differences above, we can determine if removing  $T_{i_j}(n)$  tokens give our opponent a losing position. For us to do this, we must first introduce the *gap-vector*.

**Definition 15.** Let  $n = F_{i_r} + F_{i_{r-1}} + \dots + F_{i_1}$ . We define the *gap-vector* of  $n$  to be  $G(n) = (a_r, a_{r-1}, \dots, a_2; a_1)$  where  $a_r = i_r - i_{r-1}$ ,  $a_{r-1} = i_{r-1} - i_{r-2}$ ,  $\dots$ ,  $a_2 = i_2 - i_1$ , and  $a_1 = i_1$ . We also define  $|G(n)| = r$ , where  $r$  is the number of summands in the Zeckendorf representation of  $n$ .

**Example 16.** Let  $n = 129 = F_{11} + F_9 + F_5 + F_2$ . Then,

$$G(129) = (11 - 9, 9 - 5, 5 - 2; 2) = (2, 4, 3; 2) \quad \text{and} \quad |G(129)| = 4.$$

The gap-vector of  $n$  shows the difference of the subscripts of the consecutive\* Fibonacci numbers in the Zeckendorff representation of  $n$  (again, consecutive\* refers to the subscripts  $i_j, i_{j+1}$  for some  $j \in \mathbb{N}$ ). The last coordinate of the gap-vector is the subscript of the smallest Fibonacci number present in the Zeckendorff representation of  $n$ . It follows that we can reconstruct  $n$  by using

**Example 17.** Let  $G(n) = (2, 4, 3; 2)$ . Then,  $F_2$  is the first Fibonacci number in the representation of  $n$ . From here, we can build the rest of the numbers:  $2+3 = 5$ , so  $F_5$  is the next number;  $5+4 = 9$ , so  $F_9$  is the third number; and  $9+2=11$ , so  $F_{11}$  is the last number in the representation of  $n$ . Hence,  $n = F_{11} + F_9 + F_5 + F_2 = 129$ .

It is worth mentioning that by the ZRT each  $a_j \geq 2$  for  $j \in \mathbb{N}$ . We now begin to examine which tails provide winning moves. Consider  $p = T_j(n)$  for some  $n, j \in \mathbb{N}$ . We will classify exactly when  $T_j(n)$  is a winning move and hence leaves the opponent the losing position  $(n - p, 2p)$ .

*Notational remark.* For the following lemmas, we introduce the symbol  $(k : 2)$  such that  $(k : 2) \in \{2, 3\}$  where  $(k : 2) \equiv k \pmod 2$ . Similarly,  $(k : 3) \in \{2, 3, 4\}$  where  $(k : 3) \equiv k \pmod 3$ . For example,  $F_8 + \dots + F_{8:2} = F_8 + \dots + F_2$  since  $(8 : 2) \equiv 8 \pmod 2$  and  $(8 : 2) \in \{2, 3\}$ .

For the remainder of this section, we will give a lemma and then a corollary. The lemma provides properties of particular Fibonacci series. The corollaries tie the lemma into Fibonacci Nim.

**Lemma 18.** *For  $k \geq 5$ , we have  $F_k > 2(F_{k-3} + F_{k-5} + \dots + F_{k:2})$ .*

*Proof.* For  $k = 5$  and  $k = 6$ ,

$$F_5 = 5 > 2(1) = 2(F_2),$$

$$F_6 = 8 > 4 = 2(2) = 2(F_3).$$

Suppose  $F_k > 2(F_{k-3} + F_{k-5} + \dots + F_{k:2})$ . Then by the induction hypothesis,  $2F_{k-1} + F_k > 2F_{k-1} + 2(F_{k-3} + F_{k-5} + \dots + F_{k:2}) = 2(F_{k-1} + \dots + F_{k:2})$ . But,  $F_{k+2} = F_{k+1} + F_k > 2F_{k-1} + F_k$  by Lemma 10. Hence,  $F_{k+2} > 2(F_{k-1} + \dots + F_{k:2})$ . □

**Corollary 19.** *Let  $G(n) = (a_r, a_{r-1}, \dots, a_2; a_1)$  where  $r > 1$  and  $a_j \geq 2$  for  $r \geq j > 1$ . If  $a_{q+1} \geq 3$  for some  $r > q > 1$ , then  $(n - p, 2p)$  is a losing position for  $p = T_q(n)$ .*

*Proof.* Let  $G(n) = (a_r, a_{r-1}, \dots, a_2; a_1)$  where  $r > 1$  and  $a_j \geq 2$  for  $r \geq j > 1$ . Suppose  $a_{q+1} \geq 3$  for some  $r > q > 1$  and set  $p = T_q(n)$ . Then  $i_{q+1} \geq i_q + 3$ . By Lemma 18, we have  $F_{i_{q+1}} > 2(F_{i_q} + \dots + F_{i_1}) = 2T_q(n)$ . We have,  $n - p = F_{i_r} + \dots + F_{i_{q+1}}$  by the uniqueness of Zeckendorff representations, hence  $T(n - p) = F_{i_{q+1}} > 2T_q(n) = 2p$  and  $(n - p, 2p)$  is a losing position. □

We see by the above corollary that if  $G(n) = (a_r, \dots, a_2; a_1)$  contains coordinates  $a_j \geq 2$  and some  $a_{q+1} \geq 3$  we can always remove the tail beginning with the Fibonacci number  $F_{i_q}$ . But notice, by the ZRT, every representation will have  $a_j \geq 2$  for  $r \geq j \geq 2$ . Hence, we have just shown by Corollary 19 that given some  $n = F_{i_r} + \dots + F_{i_{j+1}} + F_{i_j} + \dots + F_{i_1}$ , if  $i_{j+1} - 3 \geq i_j$ , then removing  $p = T_j(n)$  results in  $(n - p, 2p)$  being a losing position. Therefore it follows that we need only to consider when  $i_{j+1} - 2 = i_j$  to classify the remainder of winning tails.

**Lemma 20.** *For  $k \geq 8$ , we have  $F_k > 2(F_{k-2} + F_{k-6} + F_{k-8} + \dots + F_{k:2})$ .*

*Proof.* For  $k = 8$  and  $k = 9$ ,

$$F_8 = 21 > 2(8 + 1) = 2(F_6 + F_2),$$

$$F_9 = 34 > 30 = 2(13 + 2) = 2(F_7 + F_3).$$

Assume

$$F_k > 2(F_{k-2} + F_{k-6} + F_{k-8} + \cdots + F_{k:2}).$$

By the induction hypothesis we have,

$$F_{k+2} = F_{k+1} + F_k > F_{k+1} + 2F_{k-2} + 2(F_{k-6} + \cdots + F_{k:2}).$$

But,  $F_{k+1} + 2F_{k-2} = F_k + F_{k-1} + 2F_{k-3} + 2F_{k-4}$ . By Lemma 10,  $2F_{k-3} > F_{k-2}$ . Hence,  $F_{k+1} + 2F_{k-2} > F_k + F_{k-1} + F_{k-2} + 2F_{k-4} = 2(F_k + F_{k-4})$ . Therefore,  $F_{k+2} > 2(F_k + F_{k-4} + F_{k-6} + \cdots + F_{k:2})$ .  $\square$

**Corollary 21.** Let  $G(n) = (a_r, a_{r-1}, \dots, a_2; a_1)$  where  $r > 1$  and  $a_j \geq 2$  for  $r \geq j > 1$ . If  $a_q \geq 4$  and  $a_{q+1} = 2$  for some  $r \geq q > 1$ , then  $(n - p, 2p)$  is a losing position for  $p = T_q(n)$ .

*Proof.* Let  $G(n) = (a_r, a_{r-1}, \dots, a_2; a_1)$  where  $r > 1$  and  $a_j \geq 2$  for  $r \geq j > 1$ . Suppose that  $a_q \geq 4$  for some  $r \geq q > 1$  and set  $p = T_q(n)$ . Then,  $i_{q+1} - 2 = i_q \geq i_{q-1} + 4$ . By Lemma 20, we have  $F_{i_{q+1}} > 2(F_{i_q} + \cdots + F_{i_1}) = 2T_q(n)$ . We have,  $n - p = F_{i_r} + \cdots + F_{i_{q+1}}$  by the uniqueness of Zeckendorff representations, hence  $T(n - p) = F_{i_{q+1}} > 2T_q(n) = 2p$  and  $(n - p, 2p)$  is a losing position.  $\square$

We see by Corollary 21 that if  $G(n) = (a_r, \dots, a_2; a_1)$  contains coordinates  $a_j \geq 2$  and some  $a_q \geq 4$  and  $a_{q+1} = 2$ , we can always remove the tail beginning with the Fibonacci number  $F_{i_q}$ . Hence, we have just shown that given some  $n = F_{i_r} + \cdots + F_{i_{j+1}} + F_{i_j} + \cdots + F_{i_1}$ , if  $i_{q+1} - 2 = i_q \geq i_{q-1} + 4$ , then removing  $p = T_j(n)$  results in  $(n - p, 2p)$  being a *losing position*. By Corollaries 19 and 21, we have just shown that if we have  $a_{q+1} \geq 3$  or, if  $a_{q+1} = 2$  and  $a_q \geq 4$ , then  $p = T_q(n)$  is a winning move, that is,  $(n - p, 2p)$  is a losing position. Thus, what remains to examine are the cases  $a_{q+1} = 2 = a_q$  and  $a_{q+1} = 2$  and  $a_q = 3$ . We begin with the former.

**Lemma 22.** For  $k \geq 6$ , we have  $F_k \leq 2(F_{k-2} + F_{k-4})$ .

*Proof.* Let  $k = 6$ . Then,  $F_6 = 8 = 2(3 + 1) = 2(F_4 + F_2)$ . For any  $k > 6$ , we have  $F_k = 2F_{k-2} + F_{k-3}$ . By Lemma 11,  $2F_{k-4} \geq F_{k-3}$ . Hence,  $F_k = 2F_{k-2} + F_{k-3} \leq 2(F_{k-2} + F_{k-4})$ .  $\square$

**Corollary 23.** Let  $G(n) = (a_r, a_{r-1}, \dots, a_2; a_1)$  where  $r > 2$  and  $a_j \geq 2$  for  $r \geq j > 1$ . If  $a_{q+1} = 2 = a_q$  for some  $r \geq q > 1$ , then  $(n - p, 2p)$  is a winning position for  $p = T_q(n)$ .



*Proof.* Let  $G(n) = (a_r, a_{r-1}, \dots, a_2; a_1)$  where  $r > 1$  and  $a_j \geq 2$  for  $r \geq j > 1$ . Suppose  $a_{q+1} = 2 = a_q$  for some  $r \geq q > 1$  and set  $p = T_q(n)$ . Then  $i_{q+1} - 2 = i_q = i_{q-1} + 2$ . By Lemma 22, we have  $F_{i_{q+1}} \leq 2(F_{i_q} + F_{i_{q-1}}) \leq 2(F_{i_q} + \dots + F_{i_1}) = 2T_q(n)$ . We have,  $n - p = F_{i_r} + \dots + F_{i_{q+1}}$  by the uniqueness of Zeckendorff representations, but  $T(n - p) = F_{i_{q+1}} \leq 2T_q(n) = 2p$ . Thus,  $(n - p, 2p)$  is a winning position.  $\square$

We are now left with the case  $a_{q+1} = 2$  and  $a_q = 3$ . It turns out, this case is slightly more complicated than the previous cases. We will show that given  $G(n) = (a_r, a_{r-1}, \dots, a_2; a_1)$  where  $r > 1$  and  $a_j \geq 2$  for  $r \geq j > 1$ , if there exists some  $q$  such that every  $a_k \geq 3$  for  $r > q \geq k > 1$ , then  $T_q(n)$  for  $r > q > 1$  is a winning move. If however, we have some  $a_k = 2$  for  $q \geq k > 1$ , then  $T_q(n)$  for  $r > q > 1$  is a losing move. We begin with the former.

**Lemma 24.** *For  $k \geq 10$ ,  $F_k - 2(F_{k-2} + F_{k-5} + F_{k-8} + \dots + F_{k:3}) = q$  where  $q \in \{1, 2\}$ .*

*Proof.* We prove the lemma in cases for  $F_{k:3}$ . Specifically for some  $m \in \mathbb{N}$  and  $m \geq 3$ ,  $F_{k:3} = F_2$  when  $k = 3m + 1$  since  $3m + 1 - (2 + 3(m - 1)) = 2$  and  $F_{k:3} = F_3$  when  $k = 3m + 2$  since  $3m + 2 - (2 + 3(m - 1)) = 3$ .  $F_{k:3} = F_4$  when  $k = 3m$  since  $3m - (2 + 3(m - 2)) = 4$ . Note, if we have  $3m - (2 + 3(m - 1)) = 1$ , we will not have a valid Zeckendorff representation, hence we must reduce our multiple by one, which yields  $3(m - 2)$  above.

*Case 1.* Let  $F_{k:3} = F_2$  and let  $m = 3$  so that  $k = 3m + 1 = 10$ . In this case,  $F_{10} - 2(F_8 + F_5 + F_2) = 55 - 2(21 + 5 + 1) = 1$ . Let  $m > 3$  so that  $k > 10$  and assume that  $F_{3m+1} - 2(F_{3m-1} + F_{3m-4} + F_{3m-7} + \dots + F_5 + F_2) = 1$ . Then,  $F_{3m+1} + 2F_{3m+2} - 2F_{3m+2} - 2(F_{3m-1} + F_{3m-4} + \dots + F_5 + F_2) = 1$  by the inductive hypothesis. But,  $F_{3(m+1)+1} = F_{3m+4} = F_{3m+3} + F_{3m+2} = 2F_{3m+2} + F_{3m+1}$ . Hence,  $F_{3m+4} - 2(F_{3m+2} + F_{3m-1} + F_{3m-4} + \dots + F_5 + F_2) = 1$ .

*Case 2.* Now suppose that  $F_{k:3} = F_3$  and let  $m = 3$  so that  $k = 11$ . In this case,  $F_{11} - 2(F_9 + F_6 + F_3) = 89 - 2(34 + 8 + 2) = 1$ . Let  $m > 3$  so that  $k > 11$  and assume that  $F_{3m+2} - 2(F_{3m} + F_{3m-3} + F_{3m-6} + \dots + F_6 + F_3) = 1$ . Then  $F_{3m+2} + 2F_{3m+3} - 2F_{3m+3} - 2(F_{3m} + F_{3m-3} + \dots + F_6 + F_3) = 1$  by the inductive hypothesis. But,  $F_{3(m+1)+2} = F_{3m+5} = F_{3m+4} + F_{3m+3} = 2F_{3m+3} + F_{3m+2}$ . Hence,  $F_{3m+5} - 2(F_{3m+3} + F_{3m} + F_{3m-3} + \dots + F_6 + F_3) = 1$ .

*Case 3.* Finally, let  $F_{k:3} = F_4$  and let  $m = 4$  so that  $k = 12$ . Here we have  $F_{12} - 2(F_{10} + F_7 + F_4) = 144 - 2(55 + 13 + 3) = 2$ . Let  $m > 4$  so that  $k > 12$  and assume that  $F_{3m} - 2(F_{3m-2} + F_{3m-5} + F_{3m-8} + \dots + F_7 + F_4) = 2$ . Then,  $F_{3m} + 2F_{3m+1} - 2F_{3m+1} - 2(F_{3m-2} + F_{3m-5} + \dots + F_7 + F_4) = 2$  by the inductive hypothesis. But,  $F_{3(m+1)} = F_{3m+3} = F_{3m+2} + F_{3m+1} = 2F_{3m+1} + F_{3m}$ . So,  $F_{3m+3} - 2(F_{3m+1} + F_{3m-2} + F_{3m-5} + \dots + F_7 + F_4) = 2$ .

Hence, in each case we find with  $q \in \{1, 2\}$  that

$$F_k - 2(F_{k-2} + F_{k-5} + F_{k-8} + \dots + F_{k:3}) = q. \quad \square$$

**Remark 25.** It should be clear from Lemma 24 that for  $k \geq 10$ ,

$$F_k > 2((F_{k-2} + F_{k-5} + F_{k-8} + \dots + F_{k:3})).$$

**Corollary 26.** Let  $G(n) = (a_r, a_{r-1}, \dots, a_2; a_1)$  where  $r > 2$  and  $a_j \geq 2$  for  $r \geq j > 1$ . If  $a_{q+1} = 2$  and  $a_j \geq 3$  for  $q \geq j \geq 1$ , then  $(n - p, 2p)$  is a losing position for  $p = T_q(n)$ .

*Proof.* Let  $G(n) = (a_r, a_{r-1}, \dots, a_2; a_1)$  where  $r > 2$  and  $a_j \geq 2$  for  $r \geq j > 1$ . Suppose  $a_{q+1} = 2$  and  $a_j \geq 3$  for  $q \geq j \geq 1$  and set  $p = T_q(n)$ . Then  $i_{q+1} - 2 = i_q$  and  $i_{j+1} - 3 \geq i_j$  for every  $q > j \geq 1$ . By Lemma 24 and Remark 25, we have  $F_{i_{q+1}} > 2(F_{i_q} + \dots + F_{i_1}) = 2T_q(n)$ . We have,  $n - p = F_{i_r} + \dots + F_{i_{q+1}}$  by the uniqueness of Zeckendorff representations and  $T(n - p) = F_{i_{q+1}} > 2T_q(n) = 2p$ . Thus,  $(n - p, 2p)$  is a losing position.  $\square$

By Corollary 26, if  $G(n) = (a_r, \dots, a_2; a_1)$  contains coordinates  $a_j \geq 2$  and if for some  $a_{q+1} = 2$  we have for every  $q \geq k \geq 1$ ,  $a_k \geq 3$  then we may remove the tail beginning with the Fibonacci number  $F_{i_q}$ , that is,  $T_q(n)$ . All that remains to show is the case when at least one  $a_k = 2$ .

**Lemma 27.** For  $k \geq 6$ , we have  $F_k - (F_{k-1} + F_{k-4} + F_{k-7} + \dots + F_{k:3}) > 1$ .

*Proof.* We prove the lemma in cases for  $F_{k:3}$ . Specifically for some  $m \in \mathbb{N}$  and  $m \geq 2$ , there are three distinct possibilities: either  $F_{k:3} = F_2$  when  $k = 3m$  since  $3m - (1 + 3(m - 1)) = 2$  or  $F_{k:3} = F_3$  when  $k = 3m + 1$  since  $3m + 2 - (1 + 3(m - 1)) = 3$  or  $F_{k:3} = F_4$  when  $k = 3m + 2$  since  $3m + 2 - (1 + 3(m - 1)) = 4$ .

*Case 1.* Let  $m = 2$  so that  $k = 6$ . Then,  $F_6 - (F_5 + F_2) = 8 - (5 + 1) = 2$ . Assume  $F_k - (F_{k-1} + F_{k-4} + F_{k-7} + \dots + F_{k:3}) > 1$  for  $m > 2$ . Then, by induction hypothesis, we have  $F_{3m} + F_{3m+2} - F_{3m+2} - (F_{3m-1} + F_{3m-4} + \dots + F_2) > 1$ . But,  $F_{3m+3} = F_{3m+2} + F_{3m+1} > F_{3m+2} + F_{3m}$  and  $F_{3m+1} - F_{3m} > 2$  when  $m > 2$  by construction. Hence,  $F_{3m+3} - (F_{3m+2} + F_{3m-1} + \dots + F_2) > 1$ .

In Case 2, we replace  $k = 3m$  with  $k = 3m + 1$  and in Case 3 we replace  $k = 3m$  with  $k = 3m + 2$ . The arguments are then the same as that of Case 1.  $\square$

**Corollary 28.** Let  $G(n) = (a_r, a_{r-1}, \dots, a_2; a_1)$ . If every  $a_j = 3$  for some  $r > j > 1$  but there exists at least one  $a_q = 2$  such that  $j > q \geq 1$ , then for  $p = T_j(n)$ ,  $(n - p, 2p)$  is a winning position.

*Proof.* Let  $G(n) = (a_r, a_{r-1}, \dots, a_2; a_1)$ . Suppose that every  $a_j = 3$  for some  $r > j > 1$  except for some  $a_q = 2$  such that  $j > q \geq 1$  and set  $p = T_j(n)$ . Define  $G(n') = (b_r, b_{r-1}, \dots, r_2; r_1)$  where each  $b_j = 3$  for  $r \geq j > 1$  and  $b_1 = a_1$ . Then, by definitions 6 and 15, if  $T_q(n) = F_{i_q} + F_{i_{q-1}} + \dots + F_{i_1}$  then  $T_q(n) = F_{i_{q-1}} +$

$F_{i_{q-1}-1} + \dots + F_{i_1-1}$ . If  $i_1 - 1 = 1$ , then  $T_q(n')$  terminates with  $F_{i_2-1}$ , which will make no difference in the following argument. By Lemma 24,  $F_{i_{q+1}} - 2T_q(n') = g$  where  $g \in \{1, 2\}$ . By Lemma 27,  $F_{i_{q+1}} \geq T_q(n') + 2$ . Therefore,

$$F_{i_{q+1}} - 2T_q(n) \leq F_{i_{q+1}} - 2(T_q(n') + 2) = g - 4.$$

Since  $g \in \{1, 2\}$ ,  $g - 4 < 0$ . This immediately shows that

$$T(n - p) = F_{i_{q+1}} \leq 2T_q(n) = 2p$$

and hence  $(n - p, 2p)$  is a winning position. □

We have now fully characterized when  $T_j(n)$  is a winning move based solely on the *gap-vectors* of  $n$ . We present a table below to summarize this section’s findings. Let  $n = F_{i_r} + F_{i_{r-1}} + \dots + F_{i_1}$ . Then,  $G(n) = (a_r, a_{r-1}, \dots, a_2; a_1)$ . Recall, each  $a_j \geq 2$  by construction. Let the tail in question be  $T_j(n)$ . Then the “gaps” that surround  $F_{i_j}$  are precisely  $a_{j+1}$  and  $a_j$ . We have the following:

$a_{j+1}$	$a_j$	Further Conditions	Winning Move
$\geq 3$	$\geq 2$	None	Yes
2	$\geq 4$	None	Yes
2	2	None	No
2	3	$j \geq q \geq 1, a_q \geq 3$	Yes
2	3	$\exists q$ for $j \geq q \geq 1, a_q = 2$	No

Thus, by knowing the Zeckendorff representation of  $n$ , we may now find all possible winning moves, or moves that make  $(n - p, 2p)$  a losing position.

In Table 1, we present these results for  $n \in [1, 90] \subset \mathbb{N}$ . First, recall the first 11 Fibonacci numbers:  $F_1 = 1, F_2 = 1, F_3 = 2, F_4 = 3, F_5 = 5, F_6 = 8, F_7 = 13, F_8 = 21, F_9 = 34, F_{10} = 55, F_{11} = 89$ . The column ‘Zeck.’ gives the Zeckendorf representation in vector form, where the rightmost number is the coefficient of  $F_2$ , for example,  $17 = F_7 + F_4 + F_2 = (100101)$ . The last column lists the sum of each winning tail. Continuing with  $n = 17$ , we have  $G(17) = (3, 2; 2)$  and by the table above, we see that taking  $F_2 = 1$  and  $F_4 + F_2 = 3 + 1 = 4$  are both winning moves.

#### 4. Skilled vs unskilled players and probabilities of an unskilled win

We begin this section by noting that in order for an unskilled player to win against a skilled player, (1) the unskilled player must go first and always make a winning move, or, (2) the skilled player must start from  $n = F_k$  for some  $n, k \in \mathbb{N}$ . If not, the skilled player will immediately gain control of the game and provided the skilled player doesn’t make any mistakes, he will force a win over the nonskilled player. It is from this perspective that we discuss probabilities of an unskilled win. For

$n$	Zeck.	Moves	$n$	Zeck.	Moves	$n$	Zeck.	Moves
1	(1)	1	31	(1010010)	2; 10	61	(100001001)	1; 6
2	(10)	2	32	(1010100)	3	62	(100001010)	2; 7
3	(100)	3	33	(1010101)	1	63	(100010000)	8
4	(101)	1	34	(10000000)	34	64	(100010001)	1; 9
5	(1000)	5	35	(10000001)	1	65	(100010010)	2; 10
6	(1001)	1	36	(10000010)	2	66	(100010100)	3; 11
7	(1010)	2	37	(10000100)	3	67	(100010101)	1; 12
8	(10000)	8	38	(10000101)	1; 4	68	(100100000)	13
9	(10001)	1	39	(10001000)	5	69	(100100001)	1; 14
10	(10010)	2	40	(10001001)	1; 6	70	(100100010)	2; 15
11	(10100)	3	41	(10001010)	2; 7	71	(100100100)	3; 16
12	(10101)	1	42	(10010000)	8	72	(100100101)	1; 4; 17
13	(100000)	13	43	(10010001)	1; 9	73	(100101000)	5; 18
14	(100001)	1	44	(10010010)	2; 10	74	(100101001)	1; 6; 19
15	(100010)	2	45	(10010100)	3; 11	75	(100101010)	2; 20
16	(100100)	3	46	(10010101)	1; 12	76	(101000000)	21
17	(100101)	1; 4	47	(10100000)	13	77	(101000001)	1; 22
18	(101000)	5	48	(10100001)	1; 14	78	(101000010)	2; 23
19	(101001)	1; 6	49	(10100010)	2; 15	79	(101000100)	3; 24
20	(101010)	2	50	(10100100)	3; 16	80	(101000101)	1; 4; 25
21	(1000000)	21	51	(10100101)	1; 4	81	(101001000)	5; 26
22	(1000001)	1	52	(10101000)	5	82	(101001001)	1; 6; 27
23	(1000010)	2	53	(10101001)	1; 6	83	(101001010)	2; 7
24	(1000100)	3	54	(10101010)	2	84	(101010000)	8
25	(1000101)	1; 4	55	(100000000)	55	85	(101010001)	1; 9
26	(1001000)	5	56	(100000001)	1	86	(101010010)	2; 10
27	(1001001)	5; 6	57	(100000010)	2	87	(101010100)	3
28	(1001010)	2; 7	58	(100000100)	3	88	(101010101)	1
29	(1010000)	8	59	(100000101)	1; 4	89	(1000000000)	89
30	(1010001)	1; 9	60	(100001000)	5	90	(1000000001)	1

**Table 1.** Zeckendorff representations and winning tail sums.

the remainder of this section, we assume that the unskilled player removes tokens randomly and that the skilled player is free from making errors. Further, we commit to the following strategy for a skilled player in a losing position:

**Losing position strategy (LPS).** *If the skilled player is currently playing from a losing position, then he removes one token.*

Therefore, by Definition 9 if the skilled player is given a position  $(n, 2p')$  such that  $T(n) > 2p'$ , then set  $p = 1$  and give the opponent the position  $(n - 1, 2)$ . Hence, the unskilled player may take either one or two tokens on their next turn.

**Lemma 29.** *Let the current position be  $(n, 2)$  to the unskilled player. Then for  $p \in \{1, 2\}$ , we have  $P[(n - p, 2p) = \text{losing position}] \leq \frac{1}{2}$ .*

*Proof.* Assume the unskilled player has position  $(n, 2)$  where

$$n = F_{i_r} + F_{i_{r-1}} + \dots + F_{i_1}.$$

If  $F_{i_1} = 1 = F_2$ , then  $p = 1$  leaves  $n - 1 = F_{i_r} + F_{i_{r-1}} + \dots + F_{i_2}$  but  $T(n - 1) \geq F_4 = 3 > 2(1) = 2p$ . Now,  $p = 2$  leaves  $(n - 2, 4)$ . Since  $2 = p \neq T_j(n)$ , by Theorem 13,  $(n - 2, 4)$  is a winning position. Now suppose  $F_{i_1} = 2 = F_3$ , then the role of  $p = 1$  and  $p = 2$  are the reverse of case 1. Finally, If  $F_{i_1} = m \geq 3$ , then  $T(n) = F_{i_1} > 2$  by the ZRT. Then, by Lemma 11,  $(n - p, 2p)$  where  $p = 1$  or  $p = 2$  is a winning position. Hence, in all three instances,  $P[(n - p, 2p) = \text{losing position}] \leq \frac{1}{2}$ .  $\square$

**Lemma 30.** *Let  $n = F_{i_r} + F_{i_{r-1}} + \dots + F_{i_1}$ . Then,  $|G(n)| \leq \left\lfloor \frac{i_r}{2} \right\rfloor$ .*

*Proof.* Let  $n = F_{i_r} + F_{i_{r-1}} + \dots + F_{i_1}$  and suppose  $F_{i_r} = F_k$  from some  $k$ . Define  $n' = F_{i_{r'}} + F_{i_{r-1'}} + \dots + F_{i_1'}$  such that  $F_{i_{r'}} = F_k$  and  $G(n') = (2, 2, \dots, 2; 2)$ . Let  $k = 2m$  for some  $m \in \mathbb{N}$ . Recall, every  $a_j \geq 2$  by the ZRT. Since there are  $(2m - 2)/2 + 1 = m$  multiples of  $2 \in [2, k]$ , we have  $m = k/2 = |G(n')|$ . Suppose  $r > m$ . Then by Corollary 4 and Definition 15,  $r = |G(n)| > m$  implies that  $F_{i_r} > F_k$  which is a contradiction. Now let  $k = 2m + 1$ . Note that  $\lfloor k/2 \rfloor = m$ . Let  $n'$  be defined such that  $G(n') = (a_{r'}, a_{r-1'}, \dots, a_2'; a_1')$  where  $F_{i_{r'}} = F_k$  and each  $a_{j'} = 2$  except for some  $a_{k'} = 3$  where  $r' \geq k' \geq 1'$ . Since there are

$$\left\lfloor \frac{(2m + 1) - 2}{2} + 1 \right\rfloor = m$$

multiples of  $2 \in [2, k]$ , we have  $m = \lfloor k/2 \rfloor = |G(n')|$ . Suppose  $r > m$ . Then by Corollary 4 and Definition 15,  $r = |G(n)| > m$  implies that  $F_{i_r} > F_k$  which is a contradiction.  $\square$

Lemma 30 gives an upper bound on the number of terms in the Zeckendorff representation of some  $n$ .

**Lemma 31.** *For  $k \geq 5$ ,  $F_k \geq \frac{p^k - 0.1}{\sqrt{5}}$ , where  $p = \frac{\sqrt{5} + 1}{2}$ .*

*Proof.* The closed form of Fibonacci numbers is given by,

$$F_k = \frac{p^k - (-p)^{-k}}{\sqrt{5}}, \text{ where } p = \frac{\sqrt{5} + 1}{2}$$

(see, e.g., [Bóna 2002]). Then, we have

$$(-p)^{-5} \approx -0.09016994$$

$$(-p)^{-6} \approx 0.05572809.$$

By simple application of the derivative test from elementary calculus, we see that this is a decreasing function for all  $k \geq 5$ . Hence, we have that  $-0.1 \leq (-p)^{-k} \leq 0.1$  for all  $k \geq 5$ . Then for  $k \geq 5$ , we have

$$F_k \geq \frac{p^k - 0.1}{\sqrt{5}}. \quad \square$$

**Corollary 32.** *Let the current position be  $(n, n - 1)$  to the unskilled player where  $n \geq 5$  and  $F_{k+1} > n \geq F_k$ , then*

$$P[p = T_j(n)] \leq \frac{k\sqrt{5}}{2(p^k - 0.1)}$$

where  $1 \leq j \leq k$  and  $p$  is the unskilled player's next move.

*Proof.* If  $n = F_k$ , then  $P[p = T_j(n)] = 0$  since the only tail is  $F_k = n$  and the unskilled player may remove at most  $n - 1$  tokens. Let  $F_{k+1} > n > F_k$  so that the number of terms in the Zeckendorf representation of  $n$  is at most  $\frac{k}{2}$  terms by Lemma 30 and hence at most  $\frac{k}{2}$  possible tails. Then, since there are at least  $F_k$  possible choices for  $p$ , by Lemma 31 we have for  $1 \leq j \leq k$ ,

$$P[p = T_j(n)] = \frac{k/2}{(p^k - 0.1)/\sqrt{5}} = \frac{k\sqrt{5}}{2(p^k - 0.1)}. \quad \square$$

Corollary 32 shows that if an unskilled player begins the game where  $n \geq 5$ , then the probability that the unskilled player chooses  $p$  such that  $p$  is a winning move is less than  $\frac{2}{5}$  and by elementary calculus, the probability function  $P[p = T_j(n)]$  can be shown to be a decreasing function for  $k \geq 5$  so that as  $n$  increases, the probability that an unskilled player will choose a winning move from the beginning position (or any other of the form  $(n, n - 1)$ ) decreases exponentially. Note, if  $n = 3$  then the first player will lose and if  $n = 4$ , then only winning move the first player may take is  $p = 1$ , thus the first player has a probability of  $\frac{1}{4} < \frac{2}{5}$  of correctly choosing a tail.

We now have everything in place to state the main result of this section. This upper bound is dependent on the first move of the unskilled player however, and therefore cannot be calculated explicitly before the game begins.

**Theorem 33.** *Let  $n > p$  and  $(n - p, m)$  be the first position to the skilled player where  $m \in \{n - 1, 2p\}$ . Set  $n' = n - p$ . Then, using the LPS,*

(1) if  $n \neq F_k$  for some  $k \geq 4$  then

$$P[\text{Unskilled player wins}] \leq \frac{1}{5(2^{b-1})}, \quad \text{where } b = \left\lfloor \frac{n'}{3} \right\rfloor;$$

(2) if  $n = F_k$  for some  $k \geq 5$ , then

$$P[\text{Unskilled player wins}] \leq \frac{1}{2^b}, \quad \text{where } b = \left\lfloor \frac{n'}{3} \right\rfloor.$$

*Proof.* There are two nontrivial cases needed to prove the result.

*Case 1.*  $n \neq F_k$  for some  $n, k \in \mathbb{N}$ . If the skilled player starts, he wins every time. Thus, skilled player receives the position  $(n - p, 2p)$  where  $n - 1 \geq p \geq 1$ . By LPS, after the initial turn, the unskilled player will always receive  $(k, 2)$  for some  $k < n$  and by Lemma 29,  $P[(n - p', 2p') = \text{losing position}] \leq \frac{1}{2}$  where  $p' \in \{1, 2\}$ . Hence, at most, 3 tokens are removed after one round of play. Let  $n' = n - p$ , then there will be at least  $\lfloor n'/3 \rfloor$  rounds played from this point in the game. By Corollary 32 and repeated use of Lemma 29, we find that

$$P[\text{Unskilled player wins}] \leq \left(\frac{2}{5}\right)^{\left(\frac{1}{2^{\lfloor n'/3 \rfloor}}\right)} = \frac{1}{5(2^{b-1})}, \quad \text{where } b = \left\lfloor \frac{n'}{3} \right\rfloor.$$

*Case 2.*  $n = F_k$  for some  $n, k \in \mathbb{N}$ . By Lemma 11, removing  $p$  tokens make  $(n - p, 2p)$  a winning position. Hence, the unskilled player loses if he goes first. Now assume the skilled player begins and by LPS, takes  $1 < T(n)$  tokens. By Lemma 11,  $(n - 1, 2)$  is a winning position. Thus, this position is that of Case 1, where the unskilled player doesn't have the free move:  $(n, n - 1)$ . Hence,

$$P[\text{Unskilled player wins}] \leq \frac{1}{2^{\lfloor n'/3 \rfloor}} = \frac{1}{2^b}, \quad \text{where } b = \left\lfloor \frac{n'}{3} \right\rfloor. \quad \square$$

### 5. Final remarks

In this paper we have characterized all winning algorithms for the game Fibonacci Nim. We have shown that the known winning algorithm is just a particular case of the generalized winning algorithm. In addition, we have shown an upper bound on the probability that an unskilled player may beat a skilled player if our unskilled player guesses randomly and our skilled player plays according to our losing position strategy.

Future research may look into different losing position strategies as well as different types of unskilled players. For example, as a second losing position strategy, by taking more than one token from a losing position, we may find a tighter upper bound on the probability that the unskilled player wins. Additionally, we could introduce a semiskilled player, one whose guesses are not random but are based on some rule.

### References

- [Bóna 2002] M. Bóna, *A walk through combinatorics: An introduction to enumeration and graph theory*, World Scientific, River Edge, NJ, 2002. MR 1936456 Zbl 1043.05001
- [Hoggatt et al. 1973] V. E. Hoggatt, Jr., N. Cox, and M. Bicknell, “A primer for the Fibonacci numbers, XII”, *Fibonacci Quart.* **11**:3 (1973), 317–331. MR 48 #3859 Zbl 0274.10019
- [Holshouser et al. 2003] A. Holshouser, H. Reiter, and J. Rudzinski, “Dynamic one-pile nim”, *Fibonacci Quart.* **41**:3 (2003), 253–262. MR 2004i:05005 Zbl 1093.91013
- [Lekkerkerker 1952] C. G. Lekkerkerker, “Representation of natural numbers as a sum of Fibonacci numbers”, *Simon Stevin* **29** (1952), 190–195. MR 15,401c Zbl 0049.03101
- [Schwenk 1970] A. J. Schwenk, “Take-away games”, *Fibonacci Quart.* **8**:3 (1970), 225–234. MR 44 #1446 Zbl 0213.46402

Received: 2013-12-23      Revised: 2014-01-08      Accepted: 2014-01-24

cwallen08@gmail.com

*Department of Mathematics and Statistics,  
San Diego State University, 5500 Campanile Drive,  
San Diego, CA 92182-7720, United States*

vponomarenko@mail.sdsu.edu

*Department of Mathematics and Statistics,  
San Diego State University, 5500 Campanile Drive,  
San Diego, CA 92182-7720, United States*



## Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the *Involve* website.

**Originality.** Submission of a manuscript acknowledges that the manuscript is original and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

**Language.** Articles in *Involve* are usually in English, but articles written in other languages are welcome.

**Required items.** A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

**Format.** Authors are encouraged to use  $\LaTeX$  but submissions in other varieties of  $\TeX$ , and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

**References.** Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of Bib $\TeX$  is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

**Figures.** Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to [graphics@msp.org](mailto:graphics@msp.org) with details about how your graphics were generated.

**White space.** Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

**Proofs.** Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

# involve

2014

vol. 7

no. 6

A median estimator for three-dimensional rotation data MELISSA BINGHAM AND ZACHARY FISCHER	713
Numerical results on existence and stability of steady state solutions for the reaction-diffusion and Klein–Gordon equations MILES ARON, PETER BOWERS, NICOLE BYER, ROBERT DECKER, ASLIHAN DEMIRKAYA AND JUN HWAN RYU	723
The $h$ -vectors of PS ear-decomposable graphs NIMA IMANI, LEE JOHNSON, MCKENZIE KEELING-GARCIA, STEVEN KLEE AND CASEY PINCKNEY	743
Zero-inflated Poisson (ZIP) distribution: parameter estimation and applications to model data from natural calamities SADIE BECKETT, JOSHUA JEE, THAPELO NCUBE, SOPHIA POMPILUS, QUINTEL WASHINGTON, ANSHUMAN SINGH AND NABENDU PAL	751
On commutators of matrices over unital rings MICHAEL KAUFMAN AND LILLIAN PASLEY	769
The nonexistence of cubic Legendre multiplier sequences TAMÁS FORGÁCS, JAMES HALEY, REBECCA MENKE AND CARLEE SIMON	773
Seating rearrangements on arbitrary graphs DARYL DEFORD	787
Fibonacci Nim and a full characterization of winning moves CODY ALLEN AND VADIM PONOMARENKO	807



1944-4176(2014)7:6;1-3