involve

a journal of mathematics

msp

2012 vol. 5, no. 3

involve

msp.org/involve

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2012 is US $105/year for the electronic version, and $145/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

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

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

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

INVOLVE 5:3 (2012)
dx.doi.org/10.2140/involve.2012.5.237

Analysis of the steady states of a mathematical model for Chagas disease

Mary Clauson, Albert Harrison, Laura Shuman,
Meir Shillor and Anna Maria Spagnuolo

(Communicated by Suzanne Lenhart)

The steady states of a mathematical model for the dynamics of Chagas disease, developed by Spagnuolo et al., are studied and numerically simulated. The model consists of a system of four nonlinear ordinary differential equations for the total number of domestic carrier insects, and the infected insects, infected humans, and infected domestic animals. The equation for the vector dynamics has a growth rate of the blowfly type with a delay. In the parameter range of interest, the model has two unstable disease-free equilibria and a globally asymptotically stable (GAS) endemic equilibrium. Numerical simulations, based on the fourth-order Adams–Bashforth predictor corrector scheme for ODEs, depict the various cases.

1. Introduction

Chagas disease is wide spread in rural parts of South and Central America, where an estimated 10 million people are infected [Bilate and Cunha-Neto 2008; Cohen and Gürtler 2001; Schofield et al. 2006], and a search on the World Health Organization (WHO) web site yielded 1460 results. A summary of the state of the disease can be found at [WHO 2010]. Cases of the disease were also reported in Mexico and even a few in Southern California. The disease is transmitted by the insect *Triatoma infestans*, known as the "kissing bug", which bites the victim and then defecates around the bite wound. The parasites that cause the disease, *Trypanosoma cruzi*, which are in the bug's feces, enter the wound and spread throughout the body. The disease causes significant morbidity and eventually death, and there is no cure for the disease, after its initial stage. Currently the main way to control the spread of the disease is by insecticide spraying.

A mathematical model for the dynamics of the disease was developed in [Spagnuolo et al. 2011], where the main interest was to understand the disease spread

MSC2000: primary 92D30; secondary 34K28, 34K99, 37N25.
Keywords: Chagas disease, population dynamics, blowflies rate with delay, steady states.
This research was partially supported by the NSF-REU Grant DMS 0649099.

and how to control it by using insecticide spraying. The model consists of four nonlinear ordinary differential equations (ODEs), describing the evolution of the total numbers of the insects or vectors and of the infected vectors, infected humans, and infected household mammals, which for the sake of simplicity we call dogs. It is of the MSEIR type, but with only S (susceptibles) and I (infectives) components for the insects, humans, and dogs. The model describes a typical rural village with humans, dogs, chickens, and the vectors. Although chickens cannot be infected nor are they carriers of Chagas disease, they are a blood source for the vectors, so they contribute essentially to the disease dynamics. We refer to [Spagnuolo et al. 2011] for a detailed description of the disease and the assumptions that underlie the model. An extensive literature can be found there, in [Coffield et al. 2010], and the references therein.

This work concentrates on the steady states of the model of Spagnuolo et al. and studies their stability. The time-dependent model coefficients, with their yearly oscillations are replaced by their yearly averages. Thus, the seasonal changes in the relevant system parameters are not included here. However, they were taken into account in [Coffield et al. 2010; Spagnuolo et al. 2011].

The interest in this work lies in understanding the mathematical structure of the model without spraying, and with time-independent coefficients.

We note that a somewhat different model was studied in [Spagnuolo et al. 2012; Coffield et al. 2010], where the analysis of the steady states can be found, too. There, the growth rate in the equation for the vectors was a logistic term with delay, while in [Spagnuolo et al. 2011] and here, the so-called “blowflies” term with a delay is used ([Nicholson 1954]; see also [Wei and Li 2005] and references therein).

In addition to the stability analysis of the steady states, Section 3, we present a scheme for the numerical solutions of the model and depict two sets of simulations, Section 4. The results depict the monotone ways the system approaches the endemic steady state.

2. The model

We briefly describe the mathematical model for Chagas disease developed in [Spagnuolo et al. 2011]. It describes the population dynamics of the total numbers of: vectors (bugs), infected vectors, infected humans, and infected domestic animals (dogs) in a representative village in South America. The model was used to study the effects of periodic insecticide spraying for the control of the disease. In this work we are interested in the stability of its disease-free and endemic equilibria, so we omit the terms related to insecticide spraying.

The populations are assumed to be large enough to be governed by differential equations. The total populations of humans (N), dogs (D), and chickens (C)

are assumed to remain constant over time. We denote by $V = V(t)$ the number of carrier insects living in the houses at time t; the number of infective insects by $V_i = V_i(t)$, the number of infective humans by $N_i = N_i(t)$, and the number of infective dogs by $D_i = D_i(t)$. Each non-infected population, excluding C, is assumed to be susceptible. The rate coefficients $d_h = d_h(t)$, $d_m = d_m(t)$ and $b_i = b_i(t)$ are assumed to be periodic, with period of one year.

The *mathematical model for Chagas disease* of Spagnuolo et al., without insecticide spraying, is this:

$$V' = d_h V(t-\tau)e^{-aV(t-\tau)} - d_m V, \tag{2-1}$$

$$V_i' = b_i(V - V_i)\big(P_{NV} N_i + P_{DV} d_f D_i\big) - d_m V_i, \tag{2-2}$$

$$N_i' = b_i P_{VN}\,(N - N_i)\,V_i - \gamma_N N_i, \tag{2-3}$$

$$D_i' = b_i d_f P_{VD}(D - D_i)V_i - \gamma_D D_i, \tag{2-4}$$

$$V_i(0) = V_{i0}, \quad N_i(0) = N_{i0}, \quad D_i(0) = D_{i0},$$

$$V(t) = V_0(t), \qquad -\tau \le t \le 0. \tag{2-5}$$

Equation (2-1) describes the rate of change of the total vector population. The first term on the right-hand side is similar in form to Nicholson's *blowflies model* where the growth rate at time t (days) depends on the population size at time $t-\tau$ (days) [Gurney et al. 1980; Győri and Ladas 1991; Nicholson 1954]. However, in the Nicholson model $d_{h\tau}$ is a constant, since blowflies have only two stages of development: pupae and adult. In contrast, triatomines have six distinct stages of life: five instar stages and an adult stage. The egg hatching rate $d_{h\tau} = d_{h\tau}(t)$ at time t depends on the fraction of adult females at time $t-\tau$, as well as other factors including seasonal temperatures and blood supply. In particular, the growth term attains a maximum when the number of vectors in the village houses at time $t-\tau$ reaches the value of $1/a$. The natural death rate coefficient of the vectors is d_m. We note that (2-1) is decoupled from the other equations and can be solved separately.

Equation (2-2) models the rate of change of the number of infected vectors. The first term represents the rate of growth of the infectives. The factor $b_i(t) = b/b_{sup}$ is the biting rate of the vectors b divided by the total available blood supply $b_{sup} = N + d_f D + c_f C$, where d_f and c_f are the blood supply weights of the dogs and the chickens, respectively. The susceptible vector population is $V - V_i$, and P_{NV} and P_{DV} are the respective probabilities of a vector becoming infected from biting a human or a dog.

The rate of change in the number of infected humans, (2-3), is determined by the biting rate of infected vectors $b_i(t)V_i$ and the probability $P_{VN}(N - N_i)$ of a susceptible human catching the disease in one bite. The death rate of infective humans is $\gamma_N N_i$, where γ_N is the death rate constant, and is known to be higher than that of the susceptibles, [Rassi et al. 2009]. Equation (2-4) for infected dogs is similar, but with the addition of the factor d_f to take into account the vectors' preference to feed on dogs.

The model has time-dependent coefficients that incorporate seasonal variations in the life cycles of the vectors. The oscillatory behavior of the solutions can be found in the simulations in [Spagnuolo et al. 2011]. However, to study the steady states, which we do in the next section, we replace them with their yearly averages.

3. The steady states

We now study the steady states of the problem. To this end, we first rewrite the system using time-independent averaged coefficients. We set

$$\begin{aligned} a_1 &= d_h, & a_3 &= b_i P_{NV}, & a_5 &= b_i P_{VN}, \\ a_2 &= d_m, & a_4 &= b_i d_f P_{DV}, & a_6 &= b_i d_f P_{VD}, \end{aligned}$$

where we take each a_i, $i = 1, \dots, 6$ to be the average value, over 365 days, of its corresponding function in the *baseline* simulation case studied in [Spagnuolo et al. 2011]. These system parameters are positive constants. The definitions of the various coefficients and their values used in the baseline simulation case of the model can be found in Table 1.

To simplify the presentation, we rename the dependent variables as follows: $v = V,\ x = V_i,\ y = N_i,\ z = D_i$.

The problem in the new notation is: Find the functions $\{v, x, y, z\}$, defined on the time interval $[0, T]$, such that,

$$v' = a_1 v(t-\tau)e^{-av(t-\tau)} - a_2 v, \tag{3-1}$$

$$x' = a_3(v-x)y + a_4(v-x)z - a_2 x, \tag{3-2}$$

$$y' = a_5\,(N-y)\,x - \gamma_N y, \tag{3-3}$$

$$z' = a_6(D-z)x - \gamma_D z, \tag{3-4}$$

$$\begin{aligned} &x(0) = V_{i0}, \quad y(0) = N_{i0}, \quad z(0) = D_{i0}, \\ &v(t) = V_0(t), \qquad -\tau \le t \le 0. \end{aligned} \tag{3-5}$$

To study the long time behavior of the system (3-1)–(3-4) [Hethcote 2000; Thieme 2003], we note that the steady states or the fixed points are the solutions of the system

$$0 = a_1 \bar{v} e^{-a\bar{v}} - a_2 \bar{v}, \tag{3-6}$$

$$0 = a_3(\bar{v}-\bar{x})\bar{y} + a_4(\bar{v}-\bar{x})\bar{z} - a_2\bar{x}, \tag{3-7}$$

$$0 = a_5(N-\bar{y})\bar{x} - \gamma_N \bar{y}, \tag{3-8}$$

$$0 = a_6(D-\bar{z})\bar{x} - \gamma_D \bar{z}. \tag{3-9}$$

Symbol	Description	Units
V	total number of vectors	bugs/village
N	total number of humans	humans/village
D	total number of domestic dogs	dogs/village
C	total number of chickens	chickens/village
V_i	infected domestic triatomines	bugs/village
N_i	number of infected humans	humans/village
D_i	number of infected dogs	dogs/village
$d_{h\tau}$	egg hatching rate	1/day
d_m	death rate of bugs	1/day
τ	the delay factor	days
b	biting rate	1/day
P_{NV}	human to bug infection probability (per bite)	NA
P_{DV}	dog to bug infection probability (per bite)	NA
P_{VN}	bug to human infection probability (per bite)	NA
P_{VD}	bug to dog infection probability (per bite)	NA
d_f	human factor of one dog	NA
c_f	human factor of one chicken	NA
γ_N	mortality rate of infected humans	1/day
γ_D	mortality rate of infected dogs	1/day
a^{-1}	value of V at which growth rate the largest	bugs

Table 1. The model variables and coefficients.

The two solutions of the steady-state equation (3-6) for v are

$$\bar{v}_0 = 0 \quad \text{and} \quad \bar{v}_1 = \frac{1}{a} \log \frac{a_1}{a_2}. \tag{3-10}$$

We note that since $\bar{v}_1 > 0$, (because $a_1 > a_2$ in our setting), it follows from the results in [Wei and Li 2005] that the solution $\bar{v}_0 = 0$ is unstable. Also, when $\bar{v} = \bar{v}_0 = 0$, we have that $\bar{x} = \bar{y} = \bar{z} = 0$. So, $(0, 0, 0, 0)$ is an unstable equilibrium point of the system. This corresponds to the observation that Chagas disease is endemic in Latin America.

We turn to the steady states with a positive number $\bar{v}_1$, (3-10), of total vectors. In the baseline case we have $\bar{v}_1 \approx 31,500$. It follows from [Wei and Li 2005] that $\bar{v}_1$ is locally asymptotically stable. Moreover, it is found that the condition for intrinsic oscillations in Equation (2) of [Wei and Li 2005],

$$a_2 \tau e^{\tau a}\left(\log \frac{a_1}{a_2} - 1\right) > \frac{1}{e},$$

is not satisfied, so the delay τ does not cause any oscillations of the solution. In this case, there are two nonnegative equilibria for $\bar{x}$, $\bar{y}$, and $\bar{z}$. One is the disease-free equilibrium (0, 0, 0), and the other, an endemic state, is approximately (9239, 86, 51), as computed numerically, using the baseline parameters.

The Jacobian matrix evaluated at the disease-free equilibrium is

$$J(0,0,0) = \begin{bmatrix} -a_2 & a_3\bar{v}_1 & a_4\bar{v}_1 \\ a_5 N & -\gamma_N & 0 \\ a_6 D & 0 & -\gamma_D \end{bmatrix}.$$

This matrix has three distinct real eigenvalues, one positive and the other two negative. Therefore, (31500, 0, 0, 0) is an unstable equilibrium. In Section 4 we simulate the model in cases when the initial conditions are near (31500, 0, 0, 0).

Finally, at the endemic equilibrium ($\bar{v}_1 = 31500, 9239, 86, 51$) the Jacobian matrix at $(\bar{x}, \bar{y}, \bar{z})$ is:

$$J(\bar{x},\bar{y},\bar{z}) = \begin{bmatrix} -a_3\bar{y} - a_4\bar{z} - a_2 & a_3(\bar{v}_1 - \bar{x}) & a_4(\bar{v}_1 - \bar{x}) \\ a_5(N - \bar{y}) & -a_5\bar{x} - \gamma_N & 0 \\ a_6(D - \bar{z}) & 0 & -a_6\bar{x} - \gamma_D \end{bmatrix}.$$

A straightforward computation shows that $J(9239, 86, 51)$ has three real negative eigenvalues. Therefore, the endemic steady state (31500, 9239, 86, 51) is stable and attracting, or globally asymptotically stable (GAS). It follows from the model that under these conditions, without insecticide spraying or other interventions, the disease will persist. We note that we do not make a general statement on the conditions for the endemic steady state to be GAS, only that this is so in this case.

4. Simulations

We used the fourth-order Adams–Bashforth predictor corrector method to compute the numerical approximations of the model, equations (3-1)–(3-5). Due to the delay, a small step size of $\frac{1}{100}$ of a day was chosen. We also solved the system using other numerical schemes and they all matched our results for 1000 years of simulations. Moreover, Theorem 6.2.1 in [Bellen and Zennaro 2003, p. 156], guarantees the correctness of our numerical scheme.

The values of the parameters (with their references) used in the simulations are provided in Table 2. These were taken from [Spagnuolo et al. 2011]. The simulations were run using gfortran on a 3.0 GHz Intel Core 2 Duo CPU with Cent OS 5. A typical simulation of 100 years with 100 time steps per day (3.65×10^6 time steps) took approximately 300 seconds. It was found that very long runs, over a few hundred years (tens of millions of time steps) were computationally reproducible, which indicates that the solution algorithm was stable.

Symbol	Baseline simulation value	Reference
d_f	2.45	[Gürtler et al. 2007]
c_f	4.8	[Gürtler et al. 2007]
d_m	0.00327	Estimate from [Castanera et al. 2003]
$d_{h\tau}$	0.00613	Estimate from [Castanera et al. 2003; Gorla and Schofield 1985]
b_i	0.0000215	Estimate from [Castanera et al. 2003; Catalá 1991]
γ_N	$0.7\frac{2\ln 2}{76.12\cdot 365}+0.3\frac{\ln 2}{25\cdot 365}$	Estimate from [CIA 2009; Rassi et al. 2009]
γ_D	$\frac{\ln 2}{4\cdot 365}$	Estimated 8 years
C	100	This study
N	400	This study
D	100	This study
P_{NV}	0.03	[Cohen and Gürtler 2001]
P_{DV}	0.49	[Cohen and Gürtler 2001]
P_{VN}	0.00008	Estimate from [Cohen and Gürtler 2001]
P_{VD}	0.001	Estimate from [Cohen and Gürtler 2001]
a^{-1}	50,000	This study

Table 2. The parameters used in the baseline case.

We now present two numerical simulations of the model, with averaged coefficients, with different initial conditions, showing the convergence of the system to the endemic steady state $(\bar{v}_1, 9239, 86, 51)$. The first simulation has initial conditions that are considerably smaller than the steady state and chosen as $V(0) = 2$, $V_i(0) = 2$, $N_i(0) = 10$, and $D_i(0) = 0$. In the second example, the initial conditions were chosen to be larger than the steady state values, and the values were $V(0) = 45,000$, $V_i(0) = 10,000$, $N_i(0) = 100$, and $D_i(0) = 100$.

The results of both simulations are depicted in Figure 1. In each figure the heavy line represents the solution of the case with small initial conditions, i.e., starting near zero, and the thin line is the solution starting above the steady state. The convergence to the steady state of the total number of vectors can be seen at upper left; that of the infected vectors at upper right; infected humans at lower left; and infected dogs at lower right. It is seen clearly that each one of the populations, in both cases, converges monotonically to the steady state.

However, we stress that this monotone approach is characteristic of the system with averaged parameters. So it provides only qualitative insight at best. In the field, the parameters are affected by seasonal changes and are time dependent. This was taken into account in [Spagnuolo et al. 2011], since spraying is done once a year.

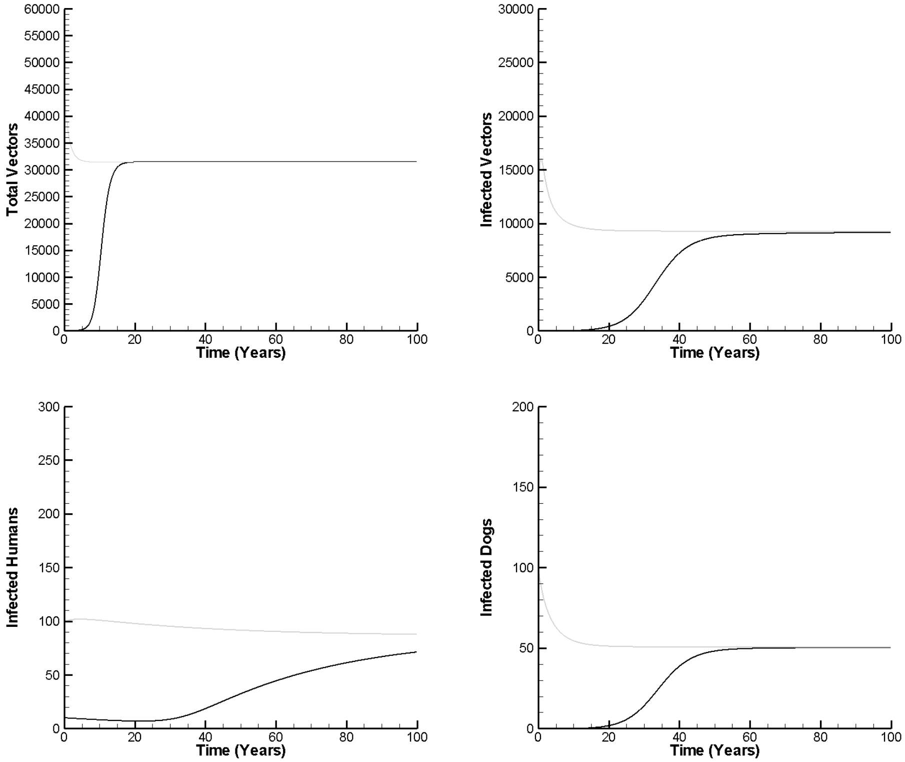


Figure 1. Convergence to the endemic state from above (thin line) and below (thick line).

5. Conclusions

A model for the dynamics of the Chagas disease, with averaged coefficients, was presented, following [Spagnuolo et al. 2012; 2011]. It consists of rate equations for the total numbers of vectors, and infected vectors, humans, and dogs (mammals). The model shows, within the conditions that seem to be observed in South America, an unstable disease-free equilibrium and a stable endemic equilibrium.

Then, our computer code was used to obtain numerical approximations of the model. In particular, we simulated the approach of the solutions to the endemic steady state. Two examples were presented, in the first one the initial conditions are below the values of the endemic equilibrium, and in the second they were above it. It was found, numerically, that the convergence to the endemic state was found to be monotone in both cases.

It may be of interest to prove that the convergence is monotone, however, the question is unresolved, yet.

Acknowledgements

The authors thank the referee for suggestions that helped improve the paper.

References

[Bellen and Zennaro 2003] A. Bellen and M. Zennaro, *Numerical methods for delay differential equations*, Clarendon, Oxford, 2003. MR 2004i:65001 Zbl 1038.65058

[Bilate and Cunha-Neto 2008] A. M. Bilate and E. Cunha-Neto, "Chagas disease cardiomyopathy: current concepts of an old disease", *Rev. Inst. Med. Trop. São Paulo* **50**:2 (2008), 67–74.

[Castanera et al. 2003] M. B. Castanera, J. P. Aparicio, and R. E. Gürtler, "A stage-structured stochastic model of the population dynamics of Triatoma infestans, the main vector of Chagas disease", *Ecol. Model.* **162** (2003), 33–53.

[Catalá 1991] S. Catalá, "The biting rate of Triatoma infestans in Argentina", *Med. Vet. Entomol.* **5**:3 (1991), 325–333.

[CIA 2009] Central Intelligence Agency, *The world factbook*, Washington, DC, 2009.

[Coffield et al. 2010] D. J. Coffield, A. M. Spagnuolo, M. Shillor, E. Mema, B. Pell, A. Pruzinsky, and A. Zetye, "A model for Chagas disease with vector consumption and congenital transmission", preprint, Oakland University, 2010.

[Cohen and Gürtler 2001] J. E. Cohen and R. E. Gürtler, "Modeling household transmission of American trypanosomiasis", *Science* **293**:5530 (2001), 694–698.

[Gorla and Schofield 1985] D. E. Gorla and C. J. Schofield, "Analysis of egg mortality in experimental populations of Triatoma infestans under natural climatic conditions in Argentina", *Bull. Soc. Vector Ecol.* **10** (1985), 107–117.

[Gurney et al. 1980] W. S. Gurney, S. P. Blythe, and R. M. Nisbet, "Nicholson's blowflies revisited", *Nature* **287** (1980), 17–21.

[Gürtler et al. 2007] R. E. Gürtler, M. C. Cecere, M. A. Lauricella, M. V. Cardinal, U. Kitron, and J. E. Cohen, "Domestic dogs and cats as sources of Trypanosoma cruzi infection in rural northwestern Argentina", *Parasitology* **134** (2007), 69–82.

[Győri and Ladas 1991] I. Győri and G. Ladas, *Oscillation theory of delay differential equations: with applications*, Clarendon, Oxford, 1991. MR 93m:34109 Zbl 0780.34048

[Hethcote 2000] H. W. Hethcote, "The mathematics of infectious diseases", *SIAM Rev.* **42**:4 (2000), 599–653. MR 2002c:92034 Zbl 0993.92033

[Nicholson 1954] A. J. Nicholson, "An outline of the dynamics of animal population", *Australian J. Zoology* **2** (1954), 9–65.

[Rassi et al. 2009] A. Rassi, Jr., A. Rassi, and J. A. Marin-Neto, "Chagas heart disease: pathophysiologic mechanisms, prognostic factors and risk stratification", *Mem. Inst. Oswaldo Cruz.* **104**:Suppl 1 (2009), 152–158.

[Schofield et al. 2006] C. J. Schofield, J. Jannin, and R. Salvatella, "The future of Chagas disease control", *Trends Parasit.* **22** (2006), 583–588.

[Spagnuolo et al. 2011] A. M. Spagnuolo, M. Shillor, and G. A. Stryker, "A model for Chagas disease with controlled spraying", *J. Biol. Dyn.* **5**:4 (2011), 299–317. MR 2012g:92205 Zbl 1219.92056

[Spagnuolo et al. 2012] A. M. Spagnuolo, M. Shillor, L. Kingsland, A. Thatcher, M. Toeniskoetter, and B. Wood, "A logistic DDE model for Chagas disease with interrupted spraying schedules", *J. Biol. Dyn.* **6**:2 (2012), 377–394.

[Thieme 2003] H. R. Thieme, *Mathematics in population biology*, Princeton University Press, Princeton, NJ, 2003. MR 2004m:92030 Zbl 1054.92042

[Wei and Li 2005] J. Wei and M. Y. Li, "Hopf bifurcation analysis in a delayed Nicholson blowflies equation", *Nonlinear Anal.* **60**:7 (2005), 1351–1367. MR 2005k:34279 Zbl 1144.34373

[WHO 2010] "Chagas disease (American trypanosomiasis)", fact sheet 340, World Health Organization, June 2010, http://www.who.int/mediacentre/factsheets/fs340/en.

Received: 2010-08-17 Revised: 2011-09-06 Accepted: 2012-01-04

maryclauson@gmail.com *Department of Biostatistics, Virginia Commonwealth University, Richmond, VA 23219, United States*

aharris2351@yahoo.com *Department of Applied Mathematics, University of Pennsylvania, Indiana 15701, United States*

lshuman@math.wsu.edu *Department of Mathematics, Washington State University, Pullman, WA 99164-3113, United States*

shillor@oakland.edu *Department of Mathematics and Statistics, Oakland University, Rochester, MI 48309-4485, United States*

spagnuol@oakland.edu *Department of Mathematics and Statistics, Oakland University, Rochester, MI 48309-4485, United States*

INVOLVE 5:3 (2012)
dx.doi.org/10.2140/involve.2012.5.247

Bounds on the artificial phase transition for perfect simulation of hard core Gibbs processes

Mark L. Huber, Elise Villella, Daniel Rozenfeld and Jason Xu

(Communicated by John C. Wierman)

Repulsive point processes arise in models where competition forces entities to be more spread apart than if placed independently. Simulation of these types of processes can be accomplished using dominated coupling from the past with a running time that depends on the intensity of the number of points. These algorithms usually exhibit what is called an artificial phase transition, where below a critical intensity the algorithm runs in finite expected time, but above the critical intensity the expected number of steps is infinite. Here the artificial phase transition is examined. In particular, an earlier lower bound on this artificial phase transition is improved by including a new type of term in the analysis. In addition, the results of computer experiments to locate the transition are presented.

1. Introduction

A spatial point process is a random collection of points in a set S. In most applications, S is a continuous space and all of the points are distinct. For instance, the locations of trees in a forest [Møller and Waagepetersen 2007] and the locations of cities in a country [Glass and Tobler 1971] can be modeled using spatial point processes.

One simple spatial point process is the Poisson point process. Suppose that S is a bounded Borel set with positive and finite Lebesgue measure. The basic Poisson point process is the outcome of the following algorithm. First choose a random number of points N according to a Poisson distribution with parameter $\lambda\mu(S)$ (so $\mathbb{P}(N=i)=\exp(-\lambda\mu(S))(\lambda\mu(S))^i/i!$ for nonnegative integers i.) Here μ is Lebesgue measure and $\lambda > 0$ is a parameter of the model. Next, choose points $X_1, \dots, X_n$ independently and uniformly from the set S. The resulting set $\{X_1, \dots, X_N\}$ is a Poisson point process.

MSC2010: primary 62M30; secondary 60G55, 60K35.

Keywords: spatial point process, dominated coupling from the past, birth-death chain.

This work was conducted as part of a summer research for undergraduates program funded through NSF grant DMS-05-48153, in cooperation with the Claremont Colleges REU.

Since the points are drawn independently, this model fails to capture situations where the locations of points are not independent. In both the forest and cities examples mentioned earlier, the points tend to be farther apart than in the independent situation since the entities involved are competing for space and resources. The points appear to act as particles with the same charge, and so they exhibit repulsion.

There are several ways to account for this repulsion. The *hard core Gibbs process* [Mase et al. 2001] is a Poisson point process conditioned on the event that none of the points lie within distance R of each other. In other words, each point is surrounded by a hard core of radius $R/2$. The cores are "hard" in the sense that the cores are not allowed to overlap. Here R is a parameter of the model.

In frequentist approaches, this model can be used to construct maximum likelihood estimators for R and λ. The values of these estimators can be approximated by methods which use random draws of the point process from the model. See, for example, [Geyer and Møller 1994; Geyer 1999; Møller and Waagepetersen 2004] for details.

In Bayesian approaches, this model (together with a prior on λ and R) can be used to build a posterior for the parameters. This posterior is quite complex, and depends on a normalizing constant (also known as partition function) that is difficult to compute exactly. The auxiliary variable method of Møller et al. [2006] can be used to create a Markov chain for these problems: this Markov chain also requires the ability to draw random variates from the model in question.

Spatial birth and death chains. Preston [1975] created a coupled pair of jump processes (X_t, Y_t) where the stationary distribution of Y_t is a Poisson point process, and the stationary distribution of X_t is the target process. In a jump process, the state stays the same until abruptly jumping to a new state (these jumps are called *events*). The time until the next jump is an exponential random variable whose rate depends only on the current state. Conditioned on this rate, the exponential is independent of all prior history of the process. For Y_t, a *birth* is an addition of a point to the process, and occurs at rate equal to $\lambda\mu(S)$. If a birth event occurs, the point added is chosen uniformly from S (again this choice is independent of the prior history of the process.) Each point when born is given a time of death that is the current time plus an exponential random variable with mean 1. This exponential is once more independent of the prior history of the process. At time of death, the point is removed from the process.

For a jump process A_t, let

$$A_{t^-} = \bigcap_{\epsilon>0} \bigcap_{t-\epsilon<t'<t} A_{t'}$$

be the state of the process immediately before time t. To use Preston's method for the hard core Gibbs process, suppose that the point v is born at time t in the Y

process. Then v is added to the X_t state if and only if it is not within distance R of a point in X_{t^-}. So births are always added to the Y process, but only sometimes to the X process in order to maintain the hard core property.

If a point $w \in Y_{t-}$ dies, at time t it is removed from the Y process. If w is also in X_{t-} it is also removed from the X process at time t. With this coupling,

$$X_{t'} \subseteq Y_{t'} \implies X_t \subseteq Y_t$$

for all $t' < t$, so the Y process is referred to as the *dominating process*.

Preston's approach yields a jump process whose limiting distribution of X_t is the hard core Gibbs process, but X_t will never exactly be in the correct distribution. Ferrari, Fernández, and Garcia [2002] developed a method for drawing samples exactly from the desired distribution using a clan of ancestors approach. In turn, Kendall and Møller [2000] developed a much faster algorithm, *dominated coupling from the past* (DCFTP), which can be used to sample from a variety of distributions that include the hard core Gibbs process.

Previous analysis showed that when using the standard Euclidean distance, the DCFTP method was provably fast when $\lambda < 1/(\pi R^2)$ [Huber 2012]. In this work we build upon this analysis, providing a wider set of conditions on λ and R for the DCFTP method to run quickly. The original argument used a term depending on the number of points in the configuration, while the new method uses the number of points as well as the area spanned by these points. This extra area term is what leads to the stronger proof. For ease of exposition we use the Euclidean metric to measure the distance between points and only operate in $\mathbb{R}^2$ throughout this work; we simply note that the same argument can easily be applied to any metric and to problems in higher dimensions.

The remainder of the work is organized as follows. Section 2 gives our new result: improved sufficient conditions on the parameters of the model for dominated coupling from the past to operate quickly. Section 3 gives computer results to complement the theoretical results of the previous section, and we close with our conclusions.

2. Bounding the running time of DCFTP

The time necessary to run DCFTP is related to the *clan of descendants* (cod) of a point v, defined as follows. For any point $v \in Y_0$, couple another point process $C_t(v)$ to Y_t as follows. Let $C_0(v) = \{v\}$. If a point w is born to Y_{t^-} at time t, add w to C_t if and only if w is within distance R of a point in C_{t^-}. If a point w' dies in the Y process at time t, and is also in the C process, remove it from C_t as well.

Then the cod of v is

$$C(v) = \bigcup_{t \geq 0} C_t(v).$$

The clan of ancestors in [Ferrari et al. 2002] is the time reversal of the cod, so they have the same size. In addition, the expected running time of DCFTP is bounded by a constant times the expected size of the cod. If there is a chance that the cod grows indefinitely, DCFTP has the same chance of taking forever to generate a sample, so the algorithm is only useful when the cod is finite with probability 1.

To bound the size of the cod, we wish to show that $\#C_t$ converges to 0 (so that $C_t = \varnothing$) with probability 1 after a finite number of births and deaths that affect the cod. In particular,

Theorem. *For $\lambda < [8/(3\sqrt{3}+4\pi)]/R^2$, the expected number of births and deaths that affect the cod is bounded above by*

$$\left(\frac{8/(3\sqrt{3}+4\pi)}{R^2} - \lambda\right)^{-1}.$$

As noted in Section 1, a similar previous result in [Huber 2012] had a constant of $1/\pi \approx 0.3183$ in front of the R^{-2} factor, while the new result has $8/(3\sqrt{3}+4\pi) \approx 0.4503$. Hence this result proves the efficacy of the DCFTP method (and mixing time of the chain) over values of λ that are 41% larger than previously known.

Avoiding boundary effects. In order to avoid having to worry about boundary effects arising from finite S, we first build another point process that dominates $C_t(v)$. As with the regular process, start with $C_0^+(v) = \{v\}$. Let $S(C_t^+(v), R)$ be all points within distance R of a point in $C_t^+(v)$. Then births in $S(C_t^+(v), R)$ will occur at rate $\lambda \cdot \mu(S(C_t^+(v), R))$. Points in C_t^+ die at rate 1. Births and deaths in S can be coupled to the births and deaths in Y_t, but there might be extra points in C_t^+ that were born outside of S. Therefore, $C_t(v) \subseteq C_t^+(v)$, and to show that $\#C_t(v)$ converges to zero, it suffices to show $\#C_t^+(v)$ converges to zero.

Useful facts. Before proving the Theorem, we show some facts that will be useful. We are only interested in how C_t^+ changes with births and deaths. Hence let t_i denote the time of the i-th event that is either a death of a point in the cod, or the proposed birth of a point within distance R of the cod. Let $D_i = C_{t_i}^+$, so D_i represents a superset of the cod after i such events have occurred. Let $\#D_i$ denote the number of points in this set.

For a configuration x, let $A(x)$ denote the Lebesgue measure of the region within distance R of at least one point in x. In particular, $A(D_i)$ is the measure of the area of the region within distance R of points in the cod. So $A(D_i)$ is proportional to the rate at which births occur that increase $\#D_i$ by 1. Our first lemma limits the average area that is added when such a birth occurs.

Lemma 1. $\mathbb{E}[A(D_{i+1}) - A(D_i) \mid$ *a birth occurs at time* $t_{i+1}] \le R^2 3\sqrt{3}/4$.

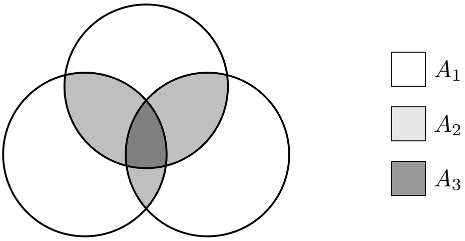


Figure 1. For circles of radius R, $3\pi R^2 = A_1 + 2A_2 + 3A_3$.

Proof. Let w be a proposed birth point. Then in order to add to the clan of descendants, w must be within distance R of a point v of D_i. The area of the new setup does not increase by πR^2, however, since only the region within R of w and not within R of v can be added area. Because w is conditioned to lie within distance R of v, the distance between centers is a random variable with density $f_r(a) = (2a/R^2) \cdot \mathbf{1}(0 \le a \le R)$.[1] Hence, the expected area added can be written as

$$\begin{aligned}\mathbb{E}[A(D_{i+1}) - A(D_i) \mid \text{birth}] &\le \int_0^R \frac{2a}{R^2}\left[\pi R^2 - 4\int_{a/2}^R \sqrt{R^2 - x^2}\, dx\right] da \\ &= R^2 3\sqrt{3}/4.\end{aligned}$$

This is an upper bound on the expected value of $A(D_{i+1}) - A(D_i)$ because w might be within distance R of other points in D_i as well, which would reduce the added area. □

The last lemma gives an upper bound on the area added when a birth occurs. The next lemma gives a lower bound on the area removed when a death occurs.

Lemma 2.

$$\mathbb{E}[A(D_{i+1}) - A(D_i) \mid \textit{ a death occurs at time } t_{i+1}] \ge [2A(D_i)/\#D_i] - \pi R^2.$$

Proof. Let A_k denote the area of the region that is within distance R of exactly k points of D_i. Then (see Figure 1)

$$\pi R^2 \# D_i = A_1 + 2A_2 + 3A_3 + \cdots + (\#D_i) A_{\#D_i},$$

and $A(D_i) = A_1 + A_2 + A_3 + \cdots + A_{\#D_i}$. Therefore

$$2A(D_i) - \pi R^2 \# D_i = A_1 - A_3 - 2A_4 - \cdots - (\#D_i - 2) A_{\#D_i} \le A_1.$$

If the points in D_i are labeled $1, 2, \ldots, \#D_i$, then $A_1 = a_1 + a_2 + \cdots + a_{\#D_1}$, where a_k is the area of the region within distance R of point i and no other points.

[1] We use $\mathbf{1}(P(a))$ for the indicator function of $P(a)$, defined as 1 if $P(a)$ is true and as 0 otherwise.

When a death occurs, every point in $\#D_i$ is equally likely to be chosen to be removed, so the average area removed is

$$\frac{1}{\#D_i}a_1 + \cdots + \frac{1}{\#D_i}a_{\#D_i} = \frac{1}{\#D_i}A_1 \geq \frac{2A(D_i)}{\#D_i} - \pi R^2. \qquad \square$$

Proof of the Theorem. For a configuration x, let $\phi(x) = A(x) + c \cdot \#x$, where $c > 0$ is a constant to be chosen later. Note that $\phi(x)$ is positive unless x is the empty configuration, in which case it equals 0. Let $\tau = \inf\{i : D_i = \varnothing\}$. Using $a \wedge b$ to denote the minimum of a and b, we shall show that $\phi(D_{i\wedge\tau}) + (i \wedge \tau)\delta$ is a supermartingale with

$$\delta = \frac{2 - \lambda R^2(3\sqrt{3}/4)}{1+\lambda}.$$

The rest of the result then follows as a consequence of the optional sampling theorem (OST). See Chapter 5 of [Durrett 2010] for a description of supermartingales and the OST.

When $i \geq \tau$, $\phi(D_{i\wedge\tau}) + (i \wedge \tau)\delta$ is a constant, and so trivially is a supermartingale.

When $i < \tau$, $\phi(D_{i+1})$ either grows when a birth occurs in the cod, or shrinks when a death occurs. First consider how $\#D_i$ changes. Births occur at rate $\lambda A(D_i)$, and deaths at rate $\#D_i$. Hence the probability that an event that changes $\#D_i$ is a birth is $A(D_i)/(A(D_i) + \#D_i)$, with the rest of the probability going towards deaths. So

$$\begin{aligned}\mathbb{E}[\#D_{i+1} - \#D_i \mid \phi(D_i)] &= \mathbb{E}\big[\mathbb{E}[\#D_{i+1} - \#D_i \mid D_i] \bigm| \phi(D_i)\big] \\ &\leq \mathbb{E}\left[\mathbf{1}(i < \tau)\left(\frac{\lambda A(D_i)}{A(D_i) + \#D_i} - \frac{\#D_i}{A(D_i) + \#D_i}\right) \middle| \phi(D_i)\right].\end{aligned}$$

(The analysis in [Huber 2012] only considered this term in ϕ, which is why the result is weaker than what is given here.)

From our first lemma, a birth increases (on average) the area covered by the cod by at most $R^2 3\sqrt{3}/4$. Our second lemma provides a lower bound on the average area removed when a death occurs. Combining these results yields

$$\mathbb{E}[A(D_{i+1}) - A(D_i) \mid \phi(D_i)] = \mathbb{E}\big[\mathbb{E}[A(D_{i+1}) - A(D_i) \mid D_i] \bigm| \phi(D_i)\big]$$

$$\leq \mathbb{E}\left[\mathbf{1}(i < \tau)\left(\frac{\lambda A(D_i)}{A(D_i) + \#D_i} R^2 \frac{3\sqrt{3}}{4} - \frac{\#D_i}{A(D_i) + \#D_i}\left(\frac{2A(D_i)}{\#D_i} - \pi R^2\right)\right) \middle| \phi(D_i)\right].$$

Note that $\mathbf{1}(i < \tau)$ is measurable with respect to $\phi(D_i)$, so bringing that out front and adding the inequalities gives

$$\begin{aligned}&\mathbb{E}[\phi(D_{i+1}) - \phi(D_i) \mid \phi(D_i)] \\ &\qquad \leq \mathbf{1}(i < \tau)\,\mathbb{E}\left[\frac{A(D_i)(\lambda((R^2 3\sqrt{3}/4) + c) - 2) + \#D_i(\pi R^2 - c)}{A(D_i) + \#D_i} \middle| \phi(D_i)\right].\end{aligned}$$

Now c can be set to

$$c = \frac{\pi R^2 + 2 - \lambda R^2(3\sqrt{3}/4)}{1+\lambda},$$

so that

$$\mathbb{E}\,[\phi(D_{i+1}) - \phi(D_i) \mid \phi(D_i)] \le \mathbf{1}(i < \tau)\mathbb{E}\left[\frac{A(D_i)(-\delta) + \#D_i(-\delta)}{A(D_i) + \#D_i} \;\middle|\; \phi(D_i)\right]$$
$$= -\delta\mathbf{1}(i < \tau).$$

Hence $\phi(D_{i\wedge\tau}) + (i \wedge \tau)\delta$ is a supermartingale. □

3. Experimental results

This theoretical result increases the known lower bound for the value of λ where the clan of descendants is finite, but this is still just a lower bound. Computer experiments can estimate this critical value of λ more precisely.

For the estimates in this section, the following protocol was used. We began a clan of descendants superset $C^+(v)$ from a single point, and recorded whether the clan died out or reached a size of 750. This was repeated 200 times, and used to estimate the probability that the clan dies out for a given value of λ. The results indicate that somewhere in [0.625, 0.626], the probability begins to drop from 1 down towards 0 (see Figure 2 for how the extinction probability changes with λ).

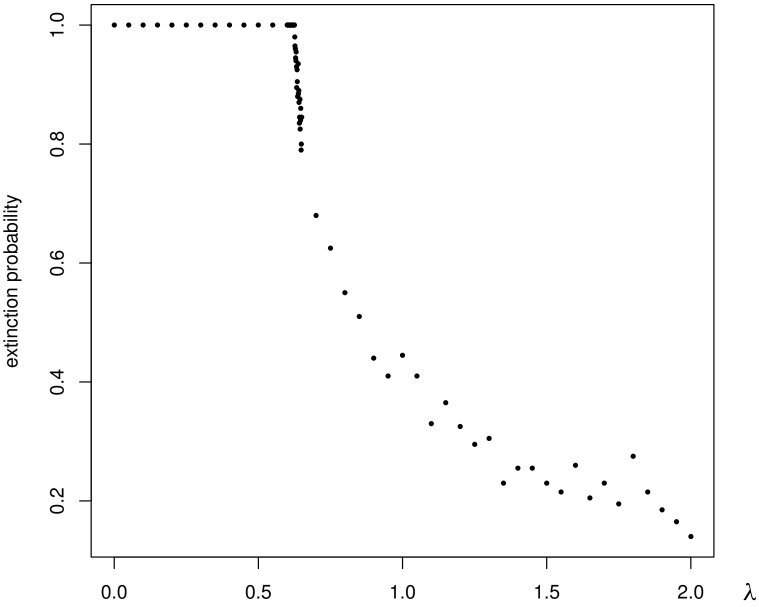


Figure 2. Estimate of extinction probability using 200 trials. The maximum cod size is 750 points.

This indicates that while the new 0.4503 theoretical result is an improvement over the old result of 0.3183, there is still work to be done to reach the true value. Increasing the ceiling size from 750 to 1500 did not alter the results within experimental error.

In short, by including a term for the area covered by the points in the potential function, a stronger theoretical lower bound on the artificial phase transition for dominated coupling from the past applied to the hard core gas model has been found. This method appears to be very general and should apply to a wide variety of repulsive processes.

References

[Durrett 2010] R. Durrett, *Probability: theory and examples*, 4th ed., Cambridge University Press, 2010. MR 2011e:60001 Zbl 1202.60001

[Ferrari et al. 2002] P. A. Ferrari, R. Fernández, and N. L. Garcia, "Perfect simulation for interacting point processes, loss networks and Ising models", *Stochastic Process. Appl.* **102**:1 (2002), 63–88. MR 2003j:60140

[Geyer 1999] C. Geyer, "Likelihood inference for spatial point processes", pp. 79–140 in *Stochastic geometry* (Toulouse, 1996), edited by O. E. Barndorff-Nielsen et al., Monogr. Statist. Appl. Probab. **80**, Chapman & Hall/CRC, Boca Raton, FL, 1999. MR 1673118

[Geyer and Møller 1994] C. J. Geyer and J. Møller, "Simulation procedures and likelihood inference for spatial point processes", *Scand. J. Statist.* **21**:4 (1994), 359–373. MR 95i:62082

[Glass and Tobler 1971] L. Glass and W. R. Tobler, "Uniform distribution of objects in a homogeneous field: Cities on a plain", *Nature* **233** (1971), 67–68.

[Huber 2012] M. L. Huber, "Spatial birth-death-swap chains", *Bernoulli* **18**:3 (2012), 1031–1041. Zbl 06064472

[Kendall and Møller 2000] W. S. Kendall and J. Møller, "Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes", *Adv. in Appl. Probab.* **32**:3 (2000), 844–865. MR 2001h:62176

[Mase et al. 2001] S. Mase, J. Møller, D. Stoyan, R. P. Waagepetersen, and G. Döge, "Packing, densities and simulated tempering for hard core Gibbs point processes", *Ann. Inst. Statist. Math.* **53**:4 (2001), 661–680. MR 2003b:60067 Zbl 1086.60512

[Møller and Waagepetersen 2004] J. Møller and R. P. Waagepetersen, *Statistical inference and simulation for spatial point processes*, Monographs on Statistics and Applied Probability **100**, Chapman & Hall/CRC, Boca Raton, FL, 2004. MR 2004h:62003 Zbl 1044.62101

[Møller and Waagepetersen 2007] J. Møller and R. P. Waagepetersen, "Modern statistics for spatial point processes", *Scand. J. Statist.* **34**:4 (2007), 643–684. MR 2009h:60091 Zbl 1157.62067

[Møller et al. 2006] J. Møller, A. N. Pettitt, R. Reeves, and K. K. Berthelsen, "An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants", *Biometrika* **93**:2 (2006), 451–458. MR 2278096

[Preston 1975] C. Preston, "Spatial birth-and-death processes", *Bull. Inst. Internat. Statist.* **46**:2 (1975), 371–391. MR 57 #14170 Zbl 0379.60082

Received: 2010-10-15 Revised: 2012-01-04 Accepted: 2012-01-13

mhuber@cmc.edu
*Department of Mathematical Sciences,
Claremont McKenna College, 850 Columbia Avenue,
Claremont, CA 91711, United States*

elisemccall@gmail.com
*Massachusetts Institute of Technology, 320 Memorial Drive,
Cambridge, MA 02139, United States*

Daniel_J_Rozenfeld@hmc.edu
*Harvey Mudd College, 340 East Foothill Boulevard,
Claremont, CA 91711, United States*

qxu@email.arizona.edu
*University of Arizona, 1021 W. Green Pebble Drive,
Tucson, AZ 85755, United States*

INVOLVE 5:3 (2012)
dx.doi.org/10.2140/involve.2012.5.257

A nonextendable Diophantine quadruple arising from a triple of Lucas numbers

A. M. S. Ramasamy and D. Saraswathy

(Communicated by Filip Saidak)

We establish that the only positive integral solutions common to the two Pell's equations $U^2 - 18V^2 = -119$ and $Z^2 - 29V^2 = -196$ are $U = 41$, $V = 10$ and $Z = 52$.

1. Introduction

Let n be a nonzero integer. We say that two integers α and β have the Diophantine property $D(n)$ if $\alpha\beta + n$ is a prefect square. A set of numbers has the property $D(n)$ if every pair of distinct elements of the set has this property. A Diophantine set S with property $D(n)$ is said to be extendable if, for some integer d, with d not belonging to S, the set $S \cup \{d\}$ is also a Diophantine set with property $D(n)$.

Sets consisting of Fibonacci numbers $\{F_m\}$ and Lucas numbers $\{L_m\}$ with the Diophantine property $D(n)$ have attracted the attention of many number theorists recently. A. Baker and H. Davenport [1969] dealt with the quadruple $\{1, 3, 8, 120\}$ with property $D(1)$ in which the first three terms are F_2, F_4 and F_6. They proved that the set cannot be extended further. V. E. Hoggatt and G. E. Bergum [1977] proved that the four numbers F_{2k}, F_{2k+2}, F_{2k+4} and $d = 4F_{2k+1}F_{2k+2}F_{2k+3}$, for $k \geq 1$, have the Diophantine property $D(1)$ and conjectured that no other integer can replace d here. The result of Baker and Davenport [1969] was an assertion of the conjecture for $k = 1$. A. Dujella [1999] proved the Hoggatt-Bergum conjecture for all positive integral values of k.

Dujella [1995] also considered Diophantine quadruples for squares of Fibonacci and Lucas numbers. In this paper we consider the Lucas numbers L_n, which are defined by $L_0 = 2$, $L_1 = 1$, $L_{n+2} = L_{n+1} + L_n$. The three Lucas numbers L_1, L_6 and L_7 have the property $D(7)$. The aim of this paper is to determine whether this set $\{1, 18, 29\}$ is extendable.

MSC2010: primary 11B39; secondary 11D09, 11B37.

Keywords: Lucas numbers, Diophantine property, recurrence relation, simultaneous Pell's equations, characteristic number, factorization, Jacobi symbol.

2. Formulation of the problem

Suppose the natural number x extends the set $S = \{1, 18, 29\}$. Then we have

$$x + 7 = V^2, \tag{1}$$

$$18x + 7 = U^2, \tag{2}$$

$$29x + 7 = Z^2, \tag{3}$$

for some integers U, V, Z. Solving (1), (2) and (3) is equivalent to solving simultaneously the two Pell's equations

$$U^2 - 18V^2 = -119, \tag{4}$$

$$Z^2 - 29V^2 = -196. \tag{5}$$

We prove that there is essentially a unique solution, so the set S can be extended by exactly one element:

Theorem. *The only positive integral solutions common to the two Pell's equations* $U^2 - 18V^2 = -119$ *and* $Z^2 - 29V^2 = -196$ *are* $U = 41$, $V = 10$ *and* $Z = 52$.

Using these values in (1) yields $x = 93$. Therefore:

Corollary. *The triple* $\{1, 18, 29\}$ *of Lucas numbers is extendable*; *the quadruple* $\{1, 18, 29, 93\}$ *has the Diophantine property* $D(7)$ *and cannot be extended further.*

3. Methodology

For the determination of the common solutions of the system of Pell's equations $3x^2 - 2 = y^2$ and $8x^2 - 7 = z^2$, Baker and Davenport [1969] gave a method based on the linear forms of logarithms of algebraic numbers. P. Kanagasabapathy and T. Ponnudurai [1975] applied quadratic reciprocity to the same system. S. P. Mohanty and A. M. S. Ramasamy [1985] introduced the concept of the characteristic number of two simultaneous Pell's equations and solved the system $U^2 - 5V^2 = -4$ and $Z^2 - 12V^2 = -11$. N. Tzanakis [2002] gave a method in for solving a system of Pell's equations using elliptic logarithms, and earlier [1993] described various methods available in the literature for finding out the common solutions of a system of Pell's equations. (For a history of numbers with the Diophantine property, one may refer to [Ramasamy 2007].)

When applying congruence methods to solve a given system of Pell's equations, the traditional approach is to work with a modulus of the form $2^\tau \cdot 3 \cdot 5$ $(\tau \geq 1)$ in the final stage of computation; see, e.g., [Kangasabapathy and Ponnudurai 1975] and [Mohanty and Ramasamy 1985]. This modulus involves only two specific odd primes, namely 3 and 5. Because of the inadequacy of such a restricted modulus for handling several problems, a method involving a general modulus was established in [Ramasamy 2006]. The present problem involves computational complexities and a new method is devised to overcome the computational difficulty by employing a result in this same reference. Taking D as a fixed natural number, one may refer to [Nagell 1951, pp. 204–212] for a theory of the general Pell's equation

$$U^2 - DV^2 = N. \tag{6}$$

We follow the conventional notations in the literature. An interesting property of Equation (6) is that its solutions may be partitioned into a certain number of disjoint classes. If m and n are two distinct integers, $U_n + V_n\sqrt{D}$ and $U_m + V_m\sqrt{D}$ belong to the same class of solutions of (6) if

$$U_n + V_n\sqrt{D} = (u + v\sqrt{D})(a + b\sqrt{D})^n, \tag{7}$$

$$U_m + V_m\sqrt{D} = (u + v\sqrt{D})(a + b\sqrt{D})^m, \tag{8}$$

where $a + b\sqrt{D}$ is the fundamental solution of Pell's equation

$$A^2 - DB^2 = 1 \tag{9}$$

and $u + v\sqrt{D}$ is the fundamental solution of (6) in the particular class. Otherwise, $U_n + V_n\sqrt{D}$ and $U_m + V_m\sqrt{D}$ belong to different classes of solutions, which are referred to as nonassociated classes (see [Nagell 1951, pp. 204–205], for example). Let $U_n + V_n\sqrt{D}$ $(n = 0, 1, 2, \dots)$ constitute a class of solutions of (6), so that we have

$$U_n + V_n\sqrt{D} = (u + v\sqrt{D})(a + b\sqrt{D})^n.$$

All the solutions of (9) with positive A and B are obtained from the formula

$$A_n + B_n\sqrt{D} = (a + b\sqrt{D})^n, \tag{10}$$

where $n = 1, 2, 3, \dots$. We have the following relations from [Mohanty and Ramasamy 1985, pp. 204–205]:

$$U_n = uA_n + DvB_n, \tag{11}$$

$$V_n = vA_n + uB_n, \tag{12}$$

$$U_{n+s} = A_sU_n + DB_sV_n, \tag{13}$$

$$V_{n+s} = B_sU_n + A_sV_n. \tag{14}$$

The sequences U_n and V_n satisfy the following recurrence relations:

$$U_{n+2} = 2aU_{n+1} - U_n, \tag{15}$$

$$V_{n+2} = 2aV_{n+1} - V_n, \tag{16}$$

$$U_{n+2s} \equiv -U_n \pmod{A_s}, \tag{17}$$

$$U_{n+2s} \equiv U_n \pmod{B_s}, \tag{18}$$

$$V_{n+2s} \equiv -V_n \pmod{A_s}, \tag{19}$$

$$V_{n+2s} \equiv V_n \pmod{B_s}. \tag{20}$$

Equations (7) and (10) imply that U_n and V_n depend on the values of A_n and B_n. In our present problem, we have $D = 18$ from (4) and therefore we have to consider the Pell equation

$$A^2 - 18B^2 = 1. \tag{21}$$

Equation (21) has the fundamental solution $A_1 = 17$, $B_1 = 4$. We check that $-67 + 16\sqrt{18}$, $-13 + 4\sqrt{18}$, $-23 + 6\sqrt{18}$ and $-41 + 10\sqrt{18}$ are the fundamental solutions of (4). Employing the condition stated for (7), we see that (4) has four nonassociated classes of solutions. Hence the general solution of (4) is given by

$$U_n + \sqrt{18}\, V_n = (-67 + 16\sqrt{18})(17 + 4\sqrt{18})^n, \tag{22}$$

$$U_n + \sqrt{18}\, V_n = (-13 + 4\sqrt{18})(17 + 4\sqrt{18})^n, \tag{23}$$

$$U_n + \sqrt{18}\, V_n = (-23 + 6\sqrt{18})(17 + 4\sqrt{18})^n, \tag{24}$$

$$U_n + \sqrt{18}\, V_n = (-41 + 10\sqrt{18})(17 + 4\sqrt{18})^n. \tag{25}$$

The solutions of (21) are provided by

$$\begin{aligned} A_0 &= 1, & A_1 &= 17, & A_{n+2} &= 34A_{n+1} - A_n, \\ B_0 &= 0, & B_1 &= 4, & B_{n+2} &= 34B_{n+1} - B_n. \end{aligned}$$

4. Solutions of the form (22)

Now, we consider the solutions of (4) given by (22), namely

$$\begin{aligned} U_0 &= -67, & U_1 &= 13, & U_{n+2} &= 34U_{n+1} - U_n, \\ V_0 &= 16, & V_1 &= 4, & V_{n+2} &= 34V_{n+1} - V_n. \end{aligned}$$

We repeatedly use the relation (19) and reason by cases.

(a) From (19) we have $V_{n+2s} \equiv -V_n \pmod{A_s}$. From this we obtain $V_{n+2} \equiv -V_n \pmod{A_1} \equiv -V_n \pmod{17}$. The sequence $V_n \pmod{17}$ is periodic with period 4. By quadratic reciprocity, we see that $n \not\equiv 0, 2 \pmod 4$. So, we are left with odd values of n only.

(b) We have $V_{n+4} \equiv -V_n \pmod{A_2} \equiv -V_n \pmod{577}$. The sequence V_n (mod 577) is periodic with period 8. We obtain $n \not\equiv 1, 3, 5, 7 \pmod 8$. Hence no solution of (4) having the form (22) satisfies (5).

5. Solutions of the form (23)

Next we consider the solutions of (4) of the form (23), namely

$$U_0=-13,\quad U_1=67,\quad U_{n+2}=34U_{n+1}-U_n,$$
$$V_0=4,\qquad V_1=16,\quad V_{n+2}=34V_{n+1}-V_n.$$

As in the previous case, one can check that no such solution can satisfy (5).

6. Solutions of the form (24)

Next we consider the solutions of (4) of the form (24), namely

$$U_0=-23,\quad U_1=41,\quad U_{n+2}=34U_{n+1}-U_n,$$
$$V_0=6,\qquad V_1=10,\quad V_{n+2}=34V_{n+1}-V_n.$$

(a) We see that $V_{n+4}\equiv -V_n \pmod{A_2}\equiv -V_n \pmod{577}$. The sequence V_n (mod 577) has period 8. By evaluating the Jacobi symbol

$$\left(\frac{V_n}{577}\right),$$

we check that $n\not\equiv 2, 3, 6, 7 \pmod 8$.

(b) We have $V_{n+6}\equiv -V_n \pmod{A_3}\equiv -V_n \pmod{1153}$. The sequence V_n (mod 1153) has period 12. It is ascertained that $n\not\equiv 8, 9 \pmod{12}$.

(c) We get $V_{n+12}\equiv -V_n \pmod{A_6}\equiv -V_n \pmod{768398401}$. On factoring, we get 768398401 = 97·577·13729. Therefore $V_{n+12}\equiv -V_n \pmod{97}$. The sequence V_n (mod 97) has period 24. We see that $n\not\equiv 4, 5, 16, 17 \pmod{24}$. Also, we have $V_{n+12}\equiv -V_n \pmod{13729}$. The sequence V_n (mod 13729) has period 24. It is seen that $n\not\equiv 0, 12 \pmod{24}$.

So far we have excluded all possibilities other than $n\equiv 1 \pmod{12}$.

(d) We obtain $V_{n+16}\equiv -V_n \pmod{A_8}\equiv -V_n \pmod{886731088897}$. We see that 886731088897 = 257·1409·2448769. Therefore $V_{n+16}\equiv -V_n \pmod{257}$. The sequence V_n (mod 257) has a period of 32. We check that $n\not\equiv 5, 9, 13, 21, 25, 29 \pmod{32}$. So we are left with $n\equiv 1 \pmod{16}$.

(e) We have $V_{n+10}\equiv -V_n \pmod{A_5}\equiv -V_n \pmod{22619537}$. We see that 22619537 = 17·241·5521. Therefore $V_{n+10}\equiv -V_n \pmod{241}$. The sequence V_n (mod 241) has period 20. We check that $n\not\equiv 5, 17 \pmod{20}$. Also $V_{n+10}\equiv -V_n \pmod{5521}$ and the sequence V_n (mod 5521) has period 20. It is seen that $n\not\equiv 9 \pmod{20}$.

(f) We get $V_{n+20}\equiv -V_n \pmod{A_{10}}\equiv -V_n \pmod{1023286908188737}$. We see that 1023286908188737 = 577·188801·9393281. Therefore $V_{n+10}\equiv -V_n$ (mod

9393281). The sequence V_n (mod 9393281) has a period of 40. We verify that $n \not\equiv 13, 33 \pmod{40}$.

The last three steps leave only the possibility $n \equiv 1 \pmod{20}$.

(g) We obtain $V_{n+14} \equiv -V_n \pmod{A_7} \equiv -V_n \pmod{26102926067}$. We see that $26102926067 = 17 \cdot 1535466241$. Therefore $V_{n+14} \equiv -V_n \pmod{1535466241}$. The sequence V_n (mod 1535466241) has period 28. We check that $n \not\equiv 5, 13, 17, 21 \pmod{28}$.

(h) We have $V_{n+28} \equiv -V_n (\mathrm{mod}\ A_{14}) \equiv -V_n \pmod{136272550150887306817}$. We see that $136272550150887306817 = 577 \cdot 209441 \cdot 11276410240481$. Therefore $V_{n+28} \equiv -V_n \pmod{209441}$. The sequence V_n (mod 209441) has period 56. We obtain $n \not\equiv 9, 25 \pmod{56}$.

Steps (d), (g) and (h) leave only the possibility $n \equiv 1 \pmod{28}$.

(i) We get $V_{n+22} \equiv -V_n \pmod{A_{11}} \equiv -V_n \pmod{34761632124320657}$. We see that $34761632124320657 = 17 \cdot 2113 \cdot 967724510017$. So $V_{n+22} \equiv -V_n \pmod{2113}$. The sequence V_n (mod 2113) has period 44. We have $n \not\equiv 9, 17, 25, 29 \pmod{44}$. Also $V_{n+22} \equiv -V_n \pmod{967724510017}$. The sequence V_n (mod 967724510017) has period 44. We get $n \not\equiv 13, 37, 41 \pmod{44}$.

(j) We have $V_{n+44} \equiv -V_n \pmod{A_{22}} \equiv -V_n \pmod{74915060494433}$. We see that $74915060494433 = 577 \cdot 129835460129$. Therefore $V_{n+44} \equiv -V_n \pmod{129835460129}$. The sequence V_n (mod 129835460129) has period 88. When $n \equiv 5, 49 \pmod{88}$, we have respectively

$$29V_n^2 - 196 \equiv 51293333469, 51271172096 \pmod{129835460129}.$$

Therefore $29V_n^2 - 196$ cannot be a square. This implies that $n \not\equiv 5, 49 \pmod{88}$. Similarly, we see that $n \not\equiv 21, 33 \pmod{88}$.

(k) We obtain $V_{n+88} \equiv -V_n \pmod{A_{44}} \equiv -V_n \pmod{2331170689}$. The sequence V_n (mod 2331170689) has a period of 176. We check that $n \not\equiv 65 \pmod{176}$.

Steps (d), (i), (j) and (k) leave only the possibility $n \equiv 1 \pmod{44}$. Consequently a solution requires $n \equiv 1 \pmod 4$, $n \equiv 1 \pmod 3$, $n \equiv 1 \pmod 5$, $n \equiv 1 \pmod 7$ and $n \equiv 1 \pmod{11}$. By the Chinese remainder theorem, then, $n \equiv 1 \pmod{2^2 \cdot 3 \cdot 5 \cdot 7 \cdot 11}$.

Now we establish that the relation $Z^2 = 29V_n^2 - 196$ is impossible for such values of n. For this purpose, we need two functions, which we now describe.

6.1. ***The functions $a(t)$ and $b(t)$.*** Throughout this subsection we keep the notation of page 259 for the solutions of the Pell equation $A^2 - DB^2 = 1$: the fundamental solution is written $a + b\sqrt{D}$ and its n-th power is $A_n + B_n\sqrt{D}$. We further consider the equation $U^2 - DV^2 = N$, singling out a class of solutions $U_n + V_n\sqrt{D} = (u + v\sqrt{D})(a + b\sqrt{D})^n$.

Definition [Mohanty and Ramasamy 1985, p. 205]. For t a natural number, define

$$\boldsymbol{a}(t) = A_{2^{t-1}} \quad \text{and} \quad \boldsymbol{b}(t) = B_{2^{t-1}}. \tag{26}$$

These functions will be used in defining a generalized characteristic number of our system of simultaneous Pell's equations. We follow [Ramasamy 2006, pp. 714–715]. We have the equalities

$$\boldsymbol{a}(t+1) = 2(\boldsymbol{a}(t))^2 - 1, \tag{27}$$

$$\boldsymbol{b}(t+1) = 2\boldsymbol{a}(t)\boldsymbol{b}(t). \tag{28}$$

Next, we have the recursion relations

$$A_n = 2aA_{n-1} - A_{n-2} \quad (n \geq 2), \tag{29}$$

$$B_n = 2aB_{n-1} - B_{n-2} \quad (n \geq 2), \tag{30}$$

which are particular cases of (15) and (16). Repeated application of these relations shows that A_n can be expressed as a polynomial in a, while B_n can be expressed as a polynomial in a and b:

$$A_n = \alpha_{n,n}a^n - \alpha_{n,n-2}a^{n-2} + \alpha_{n,n-4}a^{n-4} - \cdots, \tag{31}$$

$$B_n = \beta_{n,n}a^{n-1}b - \beta_{n,n-2}a^{n-3}b + \beta_{n,n-4}a^{n-5}b - \cdots. \tag{32}$$

Now we state a key result with reference to a system of two simultaneous Pell's equations

$$U^2 - DV^2 = N, \quad Z^2 - gV^2 = h, \tag{33}$$

where g and h are integers.

Definition and Lemma [Ramasamy 2006, Theorem 13]. *Fix odd primes $p_1 = p$, $p_2, \dots, p_s$, not necessarily distinct. Let $P = p_1 p_2 \cdots p_s$. Take $\tau \geq 1$. Set either*

(i) *$m = 2^\tau \cdot p$ and $n = i + p \cdot 2^t(2\mu + 1)$, $t \geq 1$, or*

(ii) *$m = 2^\tau \cdot P$ and $n = i + P \cdot 2^t(2\mu + 1)$, $t \geq 1$,*

where i is a fixed residue (*mod m*) *and μ is a nonnegative integer. In Case* (ii), *let $F_1, F_2, \dots$ be the polynomials contributed by the distinct primes among $p_1, p_2, \dots, p_s$ and let $G_1, G_2, \dots$ be the irreducible polynomials arising due to their various products, so that $F_1, F_2, \dots$ and $G_1, G_2, \dots$ are factors of the polynomial*

$$\beta_{P,P}D^{(P-1)/2}(\boldsymbol{b}(t+1))^{P-1} + \beta_{P,P-2}D^{(P-3)/2}(\boldsymbol{b}(t+1))^{P-3} + \cdots + \beta_{P,1}.$$

(*A prime p_i contributes a polynomial of degree $p_i - 1$. The product of two distinct primes p_i, p_j yields a factor of degree $(p_i - 1)(p_j - 1)$, and so on.*) *Let*

$$\phi := gU_i^2 - Dh \tag{34}$$

*be the **characteristic number** of the system* (33) (*for the given residue i*).

Then, for each $t \geq 1$, if at least one of the Jacobi symbols

$$\left(\frac{\phi}{(\boldsymbol{a}(t))^2 + D(\boldsymbol{b}(t))^2}\right) \quad \text{and} \quad \left(\frac{\phi}{\beta_{p,p} D^{(p-1)/2}(\boldsymbol{b}(t+1))^{p-1} + \cdots + \beta_{p,1}}\right)$$

equals -1 in Case (i), and if at least one of

$$\left(\frac{\phi}{(\boldsymbol{a}(t))^2 + D(\boldsymbol{b}(t))^2}\right), \quad \left(\frac{\phi}{F_1}\right), \quad \left(\frac{\phi}{F_2}\right), \quad \ldots, \quad \left(\frac{\phi}{G_1}\right), \quad \left(\frac{\phi}{G_2}\right), \quad \ldots$$

equals -1 in Case (ii), the system has no solution with $V = V_n$ for $n \equiv i \pmod{m}$, except possibly $V = V_i$.

6.2. ***Application of the characteristic number.*** The modulus in the present case consists of four distinct odd primes: 3, 5, 7 and 11. The characteristic number $gU_i^2 - Dh$ of the system (4), (5) for $i = 1$ is 52277; see (34). The sequence $\boldsymbol{a}(t)$ (mod 52277) is periodic with period 265 and $\boldsymbol{b}(t)$ (mod 52277) is periodic with period 530. Thus when we deal with the characteristic number of the system, we encounter computational complexities posed by the large periods of the two sequences. To overcome this difficulty, instead of working with the characteristic number directly, we consider the prime factors of the characteristic number, which are 61 and 857. The sequences $\boldsymbol{a}(t)$ (mod 61) and $\boldsymbol{b}(t)$ (mod 61) are periodic with period 5 — see Table 1 — whereas $\boldsymbol{a}(t)$ (mod 857) is periodic with period 53 and $\boldsymbol{b}(t)$ (mod 857) is periodic with period 106; moreover,

$$\boldsymbol{b}(t + 53) \equiv -\boldsymbol{b}(t) \pmod{857}. \tag{35}$$

Thus Table 2 lists only the values of $\boldsymbol{a}$ and $\boldsymbol{b}$ (mod 857) with argument up to 52. For residue calculations with respect to the factors 61 and 857, we require the values of $\boldsymbol{a}(t + 1)$ and powers of $D(\boldsymbol{b}(t + 1))^2$ modulo 61 and 857.

We take $P = 3 \cdot 5 \cdot 7 \cdot 11$ and $m = 2^\tau \cdot P$ with $\tau \geq 1$. In the notation of Case (ii) of the Definition and Lemma, we have

$$Z^2 \equiv 1185 \ (\text{mod}\ \boldsymbol{a}(t + 1) \cdot F_1 \cdots F_4 \cdot G_1 \cdots G_{11}) \tag{36}$$

where the polynomials $F_1, \ldots, G_{11}$ are illustrated in Table 3.

$t-1$	$\boldsymbol{a}(t)$	$\boldsymbol{b}(t)$
0	17	4
1	28	14
2	42	52
3	50	37
4	58	40

Table 1. Values of $\boldsymbol{a}(t)$ and $\boldsymbol{b}(t)$ (mod 61).

$t-1$	$\boldsymbol{a}(t)$	$\boldsymbol{b}(t)$	$t-1$	$\boldsymbol{a}(t)$	$\boldsymbol{b}(t)$	$t-1$	$\boldsymbol{a}(t)$	$\boldsymbol{b}(t)$	$t-1$	$\boldsymbol{a}(t)$	$\boldsymbol{b}(t)$	$t-1$	$\boldsymbol{a}(t)$	$\boldsymbol{b}(t)$
0	**17**	4	11	**652**	391	22	**563**	369	33	454	825	44	72	611
1	577	136	12	**63**	806	23	**614**	706	34	**14**	82	45	83	570
2	825	113	13	**224**	430	24	688	541	35	**391**	582	46	**65**	350
3	**333**	481	14	82	672	25	559	540	36	**669**	57	47	736	79
4	**671**	685	15	**592**	512	26	208	392	37	413	850	48	**143**	593
5	**631**	566	16	**758**	309	27	827	242	38	51	217	49	**618**	769
6	168	411	17	747	522	28	85	49	39	**59**	709	50	**260**	71
7	**742**	119	18	**203**	855	29	737	617	40	**105**	533	51	650	69
8	**739**	54	19	**145**	45	30	518	181	41	**624**	520	52	**854**	572
9	**423**	111	20	**56**	195	31	**165**	690	42	**595**	211			
10	488	493	21	**272**	415	32	**458**	595	43	167	846			

Table 2. Values of $\boldsymbol{a}(t)$ and $\boldsymbol{b}(t)$ (mod 857). For the boldface, see Note on p. 266.

$$\begin{aligned}
F_1 &= 4\bar{\boldsymbol{b}}^2+1\\
F_2 &= 16\bar{\boldsymbol{b}}^4+12\bar{\boldsymbol{b}}^2+1\\
F_3 &= 64\bar{\boldsymbol{b}}^6+80\bar{\boldsymbol{b}}^4+24\bar{\boldsymbol{b}}^2+1\\
F_4 &= 1024\bar{\boldsymbol{b}}^{10}+2304\bar{\boldsymbol{b}}^8+1792\bar{\boldsymbol{b}}^6+560\bar{\boldsymbol{b}}^4+60\bar{\boldsymbol{b}}^2+1\\
G_1 &= 256\bar{\boldsymbol{b}}^8+576\bar{\boldsymbol{b}}^6+416\bar{\boldsymbol{b}}^4+96\bar{\boldsymbol{b}}^2+1\\
G_2 &= 4096\bar{\boldsymbol{b}}^{12}+13312\bar{\boldsymbol{b}}^{10}+16384\bar{\boldsymbol{b}}^8+9344\bar{\boldsymbol{b}}^6+2368\bar{\boldsymbol{b}}^4+192\bar{\boldsymbol{b}}^2+1\\
G_3 &= 1048576\bar{\boldsymbol{b}}^{20}+5505024\bar{\boldsymbol{b}}^{18}+12320768\bar{\boldsymbol{b}}^{16}+15302656\bar{\boldsymbol{b}}^{14}+11493376\bar{\boldsymbol{b}}^{12}\\
&\quad+5326848\bar{\boldsymbol{b}}^{10}+1487104\bar{\boldsymbol{b}}^8+232256\bar{\boldsymbol{b}}^6+17440\bar{\boldsymbol{b}}^4+480\bar{\boldsymbol{b}}^2+1\\
G_4 &= 16777216\bar{\boldsymbol{b}}^{24}+104857600\bar{\boldsymbol{b}}^{22}+287309824\bar{\boldsymbol{b}}^{20}+453246976\bar{\boldsymbol{b}}^{18}+454557696\bar{\boldsymbol{b}}^{16}\\
&\quad+301907968\bar{\boldsymbol{b}}^{14}+134123520\bar{\boldsymbol{b}}^{12}+39298048\bar{\boldsymbol{b}}^{10}+7287808\bar{\boldsymbol{b}}^8+785792\bar{\boldsymbol{b}}^6\\
&\quad+40896\bar{\boldsymbol{b}}^4+576\bar{\boldsymbol{b}}^2+1
\end{aligned}$$

Table 3. Expressions for some of the polynomials in (36). We use the shorthand $\bar{\boldsymbol{b}} = \sqrt{D}\,\boldsymbol{b}(t+1)$. The polynomials $G_5, \ldots, G_{11}$ have degrees 40, 60, 48, 80, 120, 240, 480, respectively.

We still have to determine an appropriate value of t. For the application of the quadratic reciprocity law, we require the values of the polynomials modulo 4. By induction, we obtain the following results for the present problem:

$$\boldsymbol{a}(t+1) \equiv 1 \pmod 4 \quad \text{for all } t \geq 1, \tag{37}$$

$$\boldsymbol{b}(t+1) \equiv 0 \pmod 4 \quad \text{for all } t \geq 1. \tag{38}$$

We see that, for all $t \geq 1$ and i = 1, 2, 3, 4,

$$F_i, G_i \equiv 1 \pmod 4. \tag{39}$$

Considering the values of F_i and G_i modulo 857, it follows from relation (35) that F_i at $t+53$ is the same as at t, and G_i at $t+53$ is the same as at t, for all positive integers t.

6.3. ***Computations involved in the proof of the Theorem.*** With the background just provided, we are now in a position to employ the characteristic number of the present system consisting of (4) and (5). For the remaining part of our work, stagewise computation becomes necessary. The details of calculations in 9 stages required for our problem are presented in the sequel.

The characteristic number of the generalized version discussed in Section 6.1 offers several polynomials for consideration to solve a given problem, as seen from (36). First, we employ the factor $(\boldsymbol{a}(t))^2 + D(\boldsymbol{b}(t))^2$ provided by the Definition and Lemma to rule out as many possible values of t as we can.

1. *Working with* $\boldsymbol{a}(t+1)$. We consider the Jacobi symbol

$$\left(\frac{52277}{\boldsymbol{a}(t+1)}\right).$$

Using the quadratic reciprocity law and the relation (37), we evaluate this to

$$\left(\frac{61}{\boldsymbol{a}(t+1)}\right)\cdot\left(\frac{857}{\boldsymbol{a}(t+1)}\right) = \left(\frac{\boldsymbol{a}(t+1)}{61}\right)\cdot\left(\frac{\boldsymbol{a}(t+1)}{857}\right).$$

From Table 1, when $t \equiv 2, 4 \pmod 5$, we have $\boldsymbol{a}(t+1) \equiv 42, 58 \pmod{61}$, respectively; these are quadratic residues of 61. When $t \equiv 0, 1, 3 \pmod 5$, we have, respectively, $\boldsymbol{a}(t+1) \equiv 17, 28, 50 \pmod{61}$; all are quadratic nonresidues of 61.

Note. We have indicated with an asterisk in Table 2 the values of $\boldsymbol{a}(t)$ that are quadratic nonresidues of 857.

Using the fact that the product of a quadratic residue of 52277 and a nonresidue of 52277 is a nonresidue, we determine the values of t for which $\boldsymbol{a}(t+1)$ is a quadratic nonresidue of 52277. They are 1, 4, 6, 7, 9, 10, 12, 19, 22, 25, 26, 28, 30, 32, 33, 34, 38, 39, 42, 43, 45, 49, 51, 52, 55, 57, 62, 63, 64, 69, 70, 72, 74, 78, 80, 81, 83, 84, 86, 87, 89, 90, 91, 92, 94, 96, 98, 99, 100, 102, 108, 109, 114, 116, 117, 119, 120, 122, 123, 124, 127, 129, 130, 131, 133, 135, 136, 137, 142, 143, 147, 150, 151, 152, 153, 154, 159, 160, 161, 162, 164, 165, 167, 172, 173, 174, 176, 177, 179, 182, 183, 185, 186, 188, 194, 196, 199, 203, 206, 207, 209, 210, 212, 213, 217, 218, 219, 224, 226, 227, 232, 234, 236, 238, 240, 241, 244, 245, 247, 250, 252, 254, 255, 256, 262, 263, 264 (mod 265). It follows that the relation $Z^2 = 29V_n^2 - 196$ is impossible for these values of t. Therefore these values of t have to be excluded. In the sequel we consider the remaining values of t (mod 265).

t	F_1	F_2	F_3	t	F_1	F_2	F_3	t	F_1	F_2	F_3	t	F_1	F_2	F_3	t	F_1	F_2	F_3
0	**296**	**497**	**766**	11	**125**	323	**294**	22	370	149	61	33	**27**	755	**545**	44	**165**	822	**24**
1	**792**	**731**	416	12	**447**	**574**	**462**	23	518	**600**	648	34	**781**	**557**	**294**	45	**129**	**486**	**490**
2	**665**	**677**	292	13	**163**	164	166	24	**260**	**156**	177	35	480	**346**	**545**	46	**614**	529	775
3	484	778	623	14	**326**	**333**	**583**	25	**415**	**382**	**809**	36	825	**134**	**163**	47	**285**	94	32
4	404	**789**	**337**	15	658	836	72	26	796	**231**	**770**	37	101	**17**	776	48	378	142	**306**
5	335	292	79	16	**636**	**627**	**258**	27	169	**448**	575	38	117	**93**	573	49	519	**781**	240
6	**626**	**852**	**524**	17	**405**	**742**	**40**	28	616	**420**	**567**	39	209	**182**	**303**	50	**442**	**409**	775
7	**620**	**226**	35	18	289	680	658	29	**178**	**152**	**463**	40	390	800	**462**	51	**850**	41	**618**
8	**845**	**131**	**285**	19	111	433	393	30	329	587	**556**	41	**332**	2	334	52	33	264	373
9	**118**	329	468	20	**543**	**583**	**376**	31	58	**850**	386	42	**333**	668	816				
10	446	**537**	**490**	21	**268**	103	15	32	50	**835**	542	43	**143**	23	598				

Table 4. Values of F_1, F_2 and F_3 (mod 857) as functions of t (mod 53). Quadratic nonresidues of 857 are in bold.

2. *Working with* F_1. Now we consider

$$\left(\frac{52277}{F_1}\right)=\left(\frac{61}{F_1}\right)\cdot\left(\frac{857}{F_1}\right)=\left(\frac{F_1}{61}\right)\cdot\left(\frac{F_1}{857}\right),$$

in view of (39). When $t\equiv 1 \pmod 5$, we have $F_1\equiv 22 \pmod{61}$ which is a quadratic residue of 61. When $t\equiv 0, 2, 3, 4 \pmod 5$, we have $F_1\equiv 55, 38, 54, 33 \pmod{61}$; all are quadratic nonresidues of 61. As for the modulus 857, Table 4 shows the values of F_1, with the quadratic nonresidues in bold.

Consequently, we see that the relation $Z^2=29V_n^2-196$ is not true when $t\equiv 3$, 5, 11, 15, 16, 18, 21, 23, 27, 35, 37, 40, 41, 46, 48, 58, 61, 66, 68, 75, 79, 85, 88, 93, 105, 106, 110, 125, 126, 128, 132, 134, 138, 144, 145, 155, 156, 158, 163, 166, 169, 171, 178, 187, 189, 190, 195, 197, 198, 201, 208, 215, 221, 222, 226, 230, 235, 239, 242, 243, 246, 248, 249, 260 (mod 265).

3. *Working with* F_2. Next we have

$$\left(\frac{52277}{F_2}\right)=\left(\frac{61}{F_2}\right)\cdot\left(\frac{857}{F_2}\right)=\left(\frac{F_2}{61}\right)\cdot\left(\frac{F_2}{857}\right).$$

When $t\equiv 3 \pmod{61}$, we have $F_2\equiv 41 \pmod{61}$, which is a quadratic residue of 61. When $t\equiv 0, 1, 2, 4 \pmod 5$, we have, respectively, $F_2\equiv 29, 17, 17, 23 \pmod{61}$, all of which are quadratic nonresidues of 61. Further, Table 4 shows the values of F_2 modulo 857, with the quadratic nonresidues in bold.

Consequently, we see that the relation $Z^2=29V_n^2-196$ does not hold when $t\equiv$ 8, 44, 47, 53, 56, 71, 73, 95, 97, 101, 103, 104, 111, 113, 115, 118, 121, 139, 146, 149, 157, 170, 180, 181, 192, 193, 200, 202, 205, 211, 225, 228, 231, 259 (mod 265).

4. *Working with F_3.* Next we have

$$\left(\frac{52277}{F_3}\right)=\left(\frac{61}{F_3}\right)\cdot\left(\frac{857}{F_3}\right)=\left(\frac{F_3}{61}\right)\cdot\left(\frac{F_3}{857}\right).$$

When $t\equiv 1,2,3 \pmod 5$, we have respectively $F_3\equiv 3,15,5 \pmod{61}$, all of which are quadratic residues of 61. When $t\equiv 0,4 \pmod 5$, we have respectively $F_3\equiv 44,17 \pmod{61}$ both of which are quadratic nonresidues of 61. Further, Table 4 shows the values of F_3 modulo 857, with the quadratic nonresidues in bold.

As a result, the relation $Z^2=29V_n^2-196$ does not hold when $t\equiv$ 17, 24, 36, 50, 54, 60, 67, 82, 112, 141, 214, 216, 223, 237, 251, 257 (mod 265).

5. *Working with F_4.* Next we have

$$\left(\frac{52277}{F_4}\right)=\left(\frac{61}{F_4}\right)\cdot\left(\frac{857}{F_4}\right)=\left(\frac{F_4}{61}\right)\cdot\left(\frac{F_4}{857}\right).$$

When $t\equiv 0,1,4 \pmod 5$, we have respectively $F_4\equiv 42,34,4 \pmod{61}$, all of which are quadratic residues of 61. When $t\equiv 2,3 \pmod 5$, we have respectively $F_4\equiv 55,55 \pmod{61}$. It is checked that 55 is a quadratic nonresidue of 61. The relevant values modulo 857 are as follows (bold indicates quadratic nonresidues):

t (mod 53)	2	9	13	14	42	16	23	32
F_4 (mod 857)	407	827	762	**792**	619	**415**	**437**	**557**

Consequently, the relation $Z^2=29V_n^2-196$ does not hold when $t\equiv$ 2, 13, 14, 76, 148, 168, 175, 191 (mod 265).

6. *Working with G_1.* Next we have

$$\left(\frac{52277}{G_1}\right)=\left(\frac{61}{G_1}\right)\cdot\left(\frac{857}{G_1}\right)=\left(\frac{G_1}{61}\right)\cdot\left(\frac{G_1}{857}\right),$$

because of (39). When $t\equiv 3 \pmod 5$, we have $G_1\equiv 46 \pmod{61}$, which is a quadratic residue of 61. When $t\equiv 0,1,2,4 \pmod 5$, we have respectively $G_1\equiv 55,51,26,28 \pmod{61}$, all of which are quadratic nonresidues of 61. The relevant values modulo 857 are as follows (again, bold indicates quadratic nonresidues):

t (mod 53)	0	6	12	17	20	21	46
G_1 (mod 857)	774	737	57	487	785	**367**	**210**

As a result, it is seen that the relation $Z^2=29V_n^2-196$ does not hold when $t\equiv 0,20,59,65,229,233,258 \pmod{265}$.

7. *Working with G_2.* Next we have

$$\left(\frac{52277}{G_2}\right)=\left(\frac{61}{G_2}\right)\cdot\left(\frac{857}{G_2}\right)=\left(\frac{G_2}{61}\right)\cdot\left(\frac{G_2}{857}\right).$$

When $t \equiv 1, 2, 4 \pmod 5$, we have respectively $G_2 \equiv 60, 34, 49 \pmod{61}$, all of which are quadratic residues of 61. When $t \equiv 0, 3 \pmod 5$, we have respectively $G_2 \equiv 23, 6 \pmod{61}$ both of which are quadratic nonresidues of 61.

When $t \equiv 8, 34, 41 \pmod{53}$, we have respectively $G_2 \equiv 72, 177, 439 \pmod{857}$, all of which are quadratic residues of 857. When $t \equiv 25, 29, 31, 45 \pmod{53}$, we have respectively $G_2 \equiv 840, 718, 507, 781 \pmod{857}$, all of which are quadratic nonresidues of 857. As a consequence, the relation $Z^2 = 29V_n^2 - 196$ does not hold when $t \equiv 29, 31, 140, 184, 204, 220, 253 \pmod{265}$.

8. *Working with G_3.* Next we have

$$\left(\frac{52277}{G_3}\right)=\left(\frac{61}{G_3}\right)\cdot\left(\frac{857}{G_3}\right)=\left(\frac{G_3}{61}\right)\cdot\left(\frac{G_3}{857}\right).$$

When $t \equiv 4 \pmod 5$, we have $G_3 \equiv 34 \pmod{61}$, which is a quadratic residue of 61. When $t \equiv 0, 1, 2, 3 \pmod 5$, we have respectively $G_3 \equiv 59, 2, 50, 21 \pmod{61}$, all of which are quadratic nonresidues of 61. When $t \equiv 49 \pmod{53}$, we have $G_3 \equiv 453 \pmod{857}$ which is a quadratic residue of 857. Hence the relation $Z^2 = 29V_n^2 - 196$ does not hold when $t \equiv 261 \pmod{265}$.

9. *Working with G_4.* Next we have

$$\left(\frac{52277}{G_4}\right)=\left(\frac{61}{G_4}\right)\cdot\left(\frac{857}{G_4}\right)=\left(\frac{G_4}{61}\right)\cdot\left(\frac{G_4}{857}\right).$$

When $t \equiv 0, 2 \pmod 5$, we have respectively $G_4 \equiv 14, 16 \pmod{61}$, both of which are quadratic residues of 61. Modulo 61, G_4 attains the same value of 31 at $t \equiv 1 \pmod 5$ and $3 \pmod 5$. When $t \equiv 4 \pmod 5$, we have $G_4 \equiv 38 \pmod{61}$. It is seen that 31 and 38 are quadratic nonresidues of 61. When $t \equiv 1, 24 \pmod{53}$, we have, respectively, $G_4 \equiv 612, 851 \pmod{857}$ both of which are quadratic nonresidues of 857. Therefore it is seen that the relation $Z^2 = 29V_n^2 - 196$ does not hold when $t \equiv 77, 107 \pmod{265}$.

***Conclusion of the argument for solutions of the form* (24).** As mentioned, the characteristic number (in the generalized version given in [Ramasamy 2006] and explained earlier in this section) places several polynomials at our disposal for solving the problem. Each polynomial can potentially exclude several values of t. Once all values of t are excluded, we need not examine the remaining polynomials. In the present case we used the polynomials $\boldsymbol{a}(t+1)$, F_1 through F_4 and G_1 through G_4 appearing in (36), and we exhausted, in the 9 steps above, all possible values of

t; that is, we showed that the relation $Z^2 = 29V_n^2 - 196$ does not hold for any value of $t \pmod{265}$. Thus we need not consider the values attained by the polynomials G_5 through G_{11} modulo 52277. This exemplifies the usefulness of the generalized characteristic number.

The conclusion is that the system of Pell's equations $U^2 - 18V^2 = -119$, $Z^2 - 29V^2 = -196$ has no solution V_n of the form (24) except possibly for $n = 1$. When $n = 1$ we obtain a common solution with $U = \pm 41$, $V = \pm 10$ and $Z = \pm 52$.

7. Solutions of the form (25)

We finally turn to the possible solutions of the form (25):

$$\begin{aligned} U_0 &= -41, & U_1 &= 23, & U_{n+2} &= 34U_{n+1} - U_n, \\ V_0 &= 10, & V_1 &= 6, & V_{n+2} &= 34V_{n+1} - V_n. \end{aligned}$$

A case-by-case calculation as in the previous section shows that the possibilities are $n \equiv 0 \pmod 4$, $n \equiv 0 \pmod 3$, $n \equiv 0 \pmod 5$, $n \equiv 0 \pmod 7$ and $n \equiv 0 \pmod{11}$. We establish that the relation $Z^2 = 29V_n^2 - 196$ is impossible in these cases as before. The characteristic number $gU_i^2 - Dh$ of the system (4) and (5) for $i = 0$ is again 52277. Since this is the same as for the previous case, the results for the solutions in Section 6 are applicable here also.

We have now taken care of all four cases (22)–(25). Putting together the conclusions of the last four sections, we see that the proof of the Theorem is complete.

Acknowledgement

The authors are thankful to the referee for suggestions towards the improvement of the paper.

References

[Baker and Davenport 1969] A. Baker and H. Davenport, "The equations $3x^2 - 2 = y^2$ and $8x^2 - 7 = z^2$", *Quart. J. Math.* **20**:1 (1969), 129–137. MR 40 #1333 Zbl 0177.06802

[Dujella 1995] A. Dujella, "Diophantine quadruples for squares of Fibonacci and Lucas numbers", *Portugal. Math.* **52**:3 (1995), 305–318. MR 97f:11015 Zbl 0844.11015

[Dujella 1999] A. Dujella, "A proof of the Hoggatt–Bergum conjecture", *Proc. Amer. Math. Soc.* **127**:7 (1999), 1999–2005. MR 99j:11015 Zbl 0937.11011

[Hoggatt and Bergum 1977] V. E. Hoggatt, Jr. and G. E. Bergum, "A problem of Fermat and the Fibonacci sequence", *Fibonacci Quart.* **15**:4 (1977), 323–330. MR 56 #15547 Zbl 0383.10007

[Kangasabapathy and Ponnudurai 1975] P. Kangasabapathy and T. Ponnudurai, "The simultaneous Diophantine equations $y^2 - 3x^2 = -2$ and $z^2 - 8x^2 = -7$", *Quart. J. Math.* **26**:1 (1975), 275–278. MR 52 #8027 Zbl 0309.10008

[Mohanty and Ramasamy 1985] S. P. Mohanty and A. M. S. Ramasamy, "The characteristic number of two simultaneous Pell's equations and its application", *Simon Stevin* **59**:2 (1985), 203–214. MR 87c:11023 Zbl 0575.10010

[Nagell 1951] T. Nagell, *Introduction to number theory*, Wiley, New York, 1951. MR 13,207b Zbl 0042.26702

[Ramasamy 2006] A. M. S. Ramasamy, "Generalized version of the characteristic number of two simultaneous Pell's equations", *Rocky Mountain J. Math.* **36**:2 (2006), 699–720. MR 2007b:11036 Zbl 1140.11017

[Ramasamy 2007] A. M. S. Ramasamy, "Sets, sequences and polynomials linked with a question of Diophantus", *Bull. Kerala Math. Assoc.* **4**:1 (2007), 109–125. MR 2008g:11046

[Tzanakis 1993] N. Tzanakis, "Explicit solution of a class of quartic Thue equations", *Acta Arith.* **64**:3 (1993), 271–283. MR 94e:11022 Zbl 0774.11014

[Tzanakis 2002] N. Tzanakis, "Effective solution of two simultaneous Pell equations by the elliptic logarithm method", *Acta Arith.* **103**:2 (2002), 119–135. MR 2003b:11022 Zbl 1003.11064

Received: 2010-10-30 Revised: 2011-10-06 Accepted: 2011-12-18

ramasamy.mat@pondiuni.edu.in
Department of Mathematics, Pondicherry University,
R.V.Nagar, Kalapet, Pondicherry 605014, India
www.pondiuni.edu.in

dsaraswathynathan@gmail.com
Department of Mathematics, Pondicherry University,
R.V.Nagar, Kalapet, Pondicherry 605014, India

INVOLVE 5:3 (2012)
dx.doi.org/10.2140/involve.2012.5.273

Alhazen’s hyperbolic billiard problem

Nathan Poirier and Michael McDaniel

(Communicated by Joseph Gallian)

Given two points inside a circle in the hyperbolic plane, we study the problem of finding an isosceles triangle inscribed in the circle so that the two points belong to distinct congruent sides. By means of a reduction to the corresponding result in Euclidean geometry, we prove that this problem cannot generally be solved with hyperbolic ruler and compass.

In his treatise on optics, written in Arabic, the scientist and mathematician Abu Ali al-Ḥasan ibn al-Haytham (965–1039) posed the problem of *finding the light path between a source and an observer by way of a fixed spherical mirror*, and gave a geometric solution for it. The problem may have been formulated much earlier, by the great Greek mathematicians of the Hellenistic era, but no surviving testimony confirms this. Thus it is fit that it carries al-Ḥasan ibn al-Haytham’s name, which was rendered as Alhazen in the Latin translation of his book — a document that played an important role in the development of modern science.

Alhazen recognized that the problem is essentially two-dimensional — the path must lie in a plane determined by the center of the sphere, the source and the observer. His solution is long, in part because he is actually studying a more general problem; see [Sabra 1982] for details. It is not a ruler-and-compass construction, as it requires an auxiliary hyperbola; in fact, apart from special cases, the problem turns out not to be solvable with ruler and compass alone, though it seems this was only proved some 50 years ago ([Elkin 1965]; see also [Riede 1989; Neumann 1998]).

In this paper, we study the hyperbolic version of Alhazen’s problem and relate it to its classical Euclidean counterpart. We use the following formulation of the problem: *Given a circle* (*in the Euclidean or the hyperbolic plane*) *and two points A and B inside it*, *construct an inscribed*, *isosceles triangle with A on one equal leg and B on the other.*

The isosceles condition is equivalent to the condition that the two legs meet at equal angles the diameter of the circle that goes through their common vertex, so this is Alhazen’s problem all right. (One can also imagine a round billiard table

MSC2010: primary 51M04, 51M10, 51M15; secondary 51M09.
Keywords: hyperbolic geometry, Alhazen.

with two points marked on the felt. A shot that goes through one of the points, hits the cushion and then goes through the other point is a solution to the problem.)

We show the following result:

Theorem. *For a given circle in the hyperbolic plane and a given circle in the Euclidean plane, there exists a bijection — indeed a homeomorphism — between Alhazen point configurations of one and those of the other, preserving in both directions the property of Alhazen constructibility with ruler and compass.*

That is, hyperbolic configurations whose Alhazen solution is constructible with (hyperbolic) ruler and compass correspond to Euclidean configurations whose Alhazen solution is constructible with (Euclidean) ruler and compass, and similarly for nonconstructible configurations.

This correspondence was unexpected to us, since hyperbolic triangles are so different from Euclidean ones — to begin with, their angles add up to arbitrary measures less than π. Generally, Euclidean ruler-and-compass constructions fail to carry over to the hyperbolic plane; even trisecting an arbitrary segment, something quite simple with Euclidean ruler and compass, cannot be done in the hyperbolic case! (See [Martin 1975, p. 483], for instance.)

Since, as already mentioned, Alhazen's problem is seldom solvable with ruler and compass in the Euclidean plane, we obtain (see Remark 2 on page 281):

Corollary. *The hyperbolic Alhazen problem is not generally solvable with ruler and compass. Indeed, for any fixed hyperbolic open disk D_H, the set of pairs of points $A, B \in D_H$ for which the Alhazen problem can be solved with ruler and compass has measure zero in $D_H \times D_H$.*

This paper is organized as follows. In Section 1 we describe the relevant models of the hyperbolic plane and spell out the hyperbolic Alhazen problem. In Section 2 we motivate the correspondence between Euclidean and hyperbolic constructions, observing it in action in a simple constructible case. The proof of the theorem is then given in Section 3.

1. Alhazen's billiard problem in hyperbolic geometry

Hyperbolic geometry has several *models*, that is, ways to name points and make calculations. We will need to use two: the Poincaré disk model and the Klein model.

The ***Poincaré model*** represents the hyperbolic plane by an open disk, that is, the set of points inside a fixed Euclidean circle.[1] This so-called *boundary circle* is not part of the hyperbolic plane. It is a "boundary" of the model only: the hyperbolic

[1] We haven't defined the hyperbolic plane. The reader new to hyperbolic geometry can imagine that it *is* the Poincaré model: the set of points (x, y) in $\mathbb{R}^2$ satisfying $x^2 + y^2 < 1$, with further features called (hyperbolic) distance, lines, angles, and so on, which we now describe.

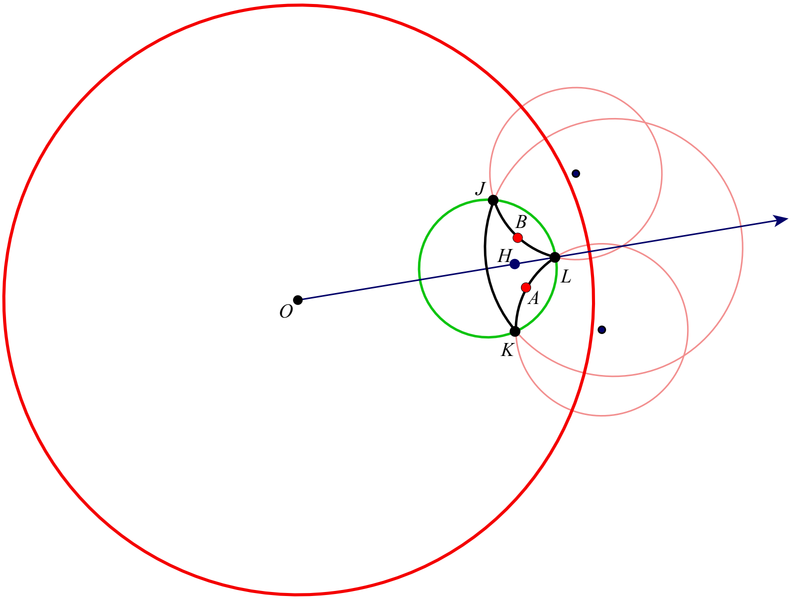


Figure 1. A hyperbolic Alhazen triangle, *JKL*. The hyperbolic plane is the interior of the disk with red boundary. The pink circles represent hyperbolic lines, to be determined in the solution of the problem, together with their intersections J, K, L. The givens of the problem consist of the small circle (on which we must place J, K, and L) and the points A and B inside it. Note the position of the center H of the given hyperbolic circle.

plane itself extends infinitely in all directions. The center of the boundary circle will be labeled O. It is not a special point in the hyperbolic plane — any point can be chosen for this honor — but it does enjoy special properties in the model.

Figure 1 illustrates the main features of the Poincaré model. The boundary circle is shown in red. Hyperbolic straight lines appear in the model either as Euclidean diameters (like the line *OH*) or as circles (in pink) orthogonal to the boundary circle — or rather, the portions of such circles inside the boundary circle. Hyperbolic circles (sets of points at a fixed hyperbolic distance from a center) appear as Euclidean circles contained in the open disk; the green circle is an example.

This is two-thirds of what we need in order to visualize the hyperbolic Alhazen problem. But how are we to recognize isosceles triangles? Hyperbolic distances cannot be discerned from appearances in the model: the formula to compute the hyperbolic distance between two points, given their coordinates in the Poincaré model, is very different from the formula giving the Euclidean distance. The ratio between the two is, roughly speaking, inversely proportional to the Euclidean distance to be boundary.

(Another manifestation of this is that the center of a hyperbolic circle does not match what appears to be the center in the model. The true center H is the point hyperbolically equidistant from the points on the circle; it can be found, for instance, as the intersection of two hyperbolic lines perpendicular to the circle. H always lies closer to the boundary than the apparent center, unless both coincide with O. In Figure 1, H is the center of the green circle.)

What saves the day is an important feature of the Poincaré model: it is *conformal*, meaning that it renders angles faithfully. The true angle between two hyperbolic lines equals the (Euclidean) angle between the circles representing the same lines in the model. Thus a hyperbolic Alhazen triangle has two equal angles in the Poincaré model, as exemplified by the triangle JKL in Figure 1. (This example is special in that the triangle's axis of symmetry, the line OH, is a diameter of the model, so the triangle also appears "isosceles", that is, symmetric, to Euclidean eyes. This would not generally be the case.)

The ***Klein model*** of the hyperbolic plane also uses a Euclidean disk to represent its points, but in this model hyperbolic lines correspond to Euclidean chords. Euclidean appearances are even more deceiving here, because hyperbolic angle measures are not the Euclidean ones visible in the model. However, the property that hyperbolic and Euclidean notions of straightness coincide in this model will be helpful.

There exists an isomorphism between the Poincaré and Klein models, based on stereographic projection, which will be the key in Section 3 to our correspondence between the hyperbolic and Euclidean Alhazen problems. To describe it, we work in (Euclidean) three-dimensional space, with both models lying on the horizontal coordinate plane. We rest a sphere on this plane, as shown in Figure 2: the radius of the sphere is half the radius of the Poincaré model, and its south pole is the center O of both models. Given a point R in the Poincaré model, we find its counterpart in the Klein model by first mapping the point onto the sphere via *stereographic projection* (central projection from the north pole); this gives a point P, somewhere on the south hemisphere. We then project P directly down onto the horizontal plane, obtaining Q; this is the counterpart of R in the Klein model.

Stereographic projection maps circles to circles and preserves orthogonality. An arc of circle orthogonal to the red boundary of the Poincaré model projects onto the sphere as a semicircle orthogonal to the equator. When projected again straight down, this gives a line segment. This confirms that in the Klein model hyperbolic lines are represented by Euclidean chords.

Finally, we observe that if a hyperbolic circle happens to be centered at O in the Poincaré model (in which case its hyperbolic and Euclidean centers coincide), it will map to a horizontal circle on the sphere, and from there down to a circle in the Klein model, again centered at O. These are the only hyperbolic circles that look like Euclidean circles in Klein model: other circles look like Euclidean ellipses.

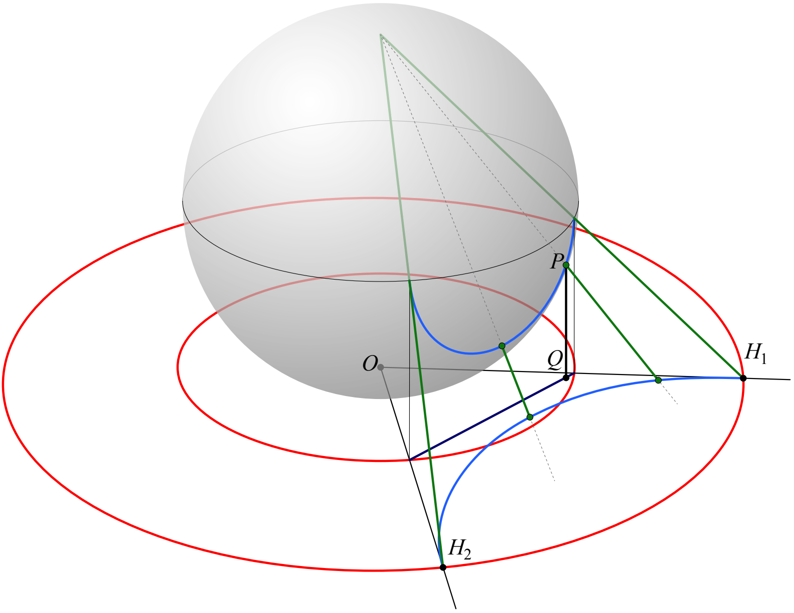


Figure 2. Correspondence between the Poincaré and Klein models.

2. A constructible example

The Euclidean Alhazen problem has an obvious solution when the given points A and B lie on a diameter of the given circle and are equidistant from the center (Figure 3, left). We simply construct the perpendicular bisector of $\overline{AB}$ — the horizontal diameter in Figure 3, right. Each of the two points where this line

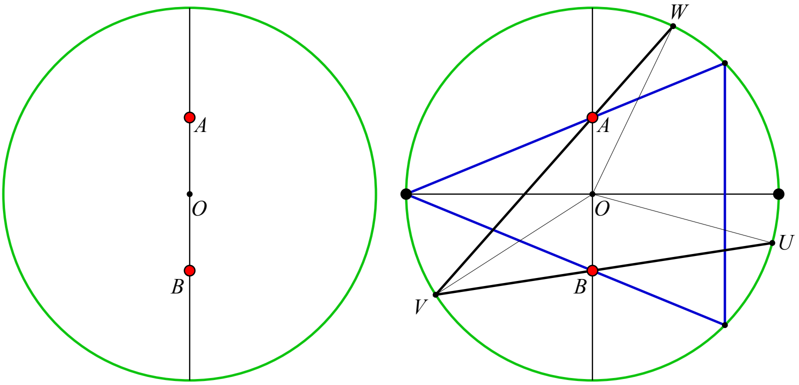


Figure 3. Solving the Alhazen problem in a special case: A and B are diametrically opposed and equidistant from the center of the circle. The triangle UVW is a solution if and only if $VO \perp AB$.

intersects the given circle provides a solution to the problem, since the angles subtended by OA and OB from these points are the same.

Moreover, these are the only solutions. To see this, take a point V on the given circle, and form the inscribed triangle whose sides lie on the lines VA and VB. If this triangle is isosceles, its median VO is also an altitude, so V lies on the perpendicular bisector of AB. (If we allow degenerate triangles then of course the line AB also provides a solution.)

For this simple situation, the reasoning in the hyperbolic case is identical. Given a hyperbolic circle of center H and two points A and B, diametrically opposed and equidistant from H, we draw the perpendicular bisector of A and B using our (hyperbolic!) compass, just as we would in the Euclidean case, and mark off its intersections with the circle, each of which provides a solution to the Alhazen problem. Seen in the Klein model, the picture would look exactly the same as Figure 3, provided we took the precaution of starting with a circle whose center coincides with the center O of the model! In the Poincaré model, with the same precaution, we'd have Figure 4.

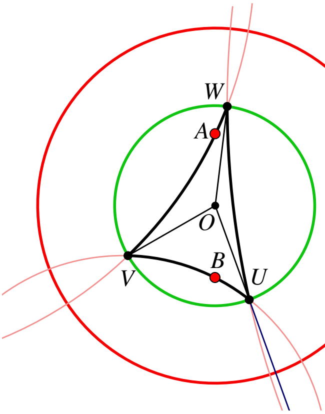


Figure 4. Special case of Alhazen problem in the Poincaré model; compare Figure 3, right. The triangle UVW is a solution if and only if $VO \perp AB$.

This leads to an important digression. With Euclidean constructions, we have physical tools (paper, ruler and compass) at our disposal, which are sufficiently accurate to help build intuition. Alas, we don't have a physical hyperbolic compass at our disposal, nor hyperbolic paper. What tools can we use to explore?

That's where the two models come in. In the Klein model, a Euclidean ruler is a proxy for a hyperbolic one, but the same cannot be said about the compass: hyperbolic circles look like ellipses in the model. What about the Poincaré model? Hyperbolic lines look like lines or circles in the model, and hyperbolic circles look like circles, so it's at least conceivable that what can be done with hyperbolic ruler and compass can also be done in the Poincaré model with Euclidean ruler and compass. And that turns out to be so:

Fact 1. *Any point in the hyperbolic plane obtained from initial data by using only* (*hyperbolic*) *ruler and compass can be obtained using ruler and compass in the Euclidean plane that supports the Poincaré model.*

This is proved by considering each building block of ruler-and-compass constructions. For instance, the problem of drawing a line through two points translates into finding a circle perpendicular to the boundary of the model and going through the given points; this *can* be done with Euclidean ruler and compass in a few steps. Drawing a (hyperbolic) circle centered at a point and going through another point translates into finding the Euclidean center of the desired circle in the model, and so on. See [Goodman-Strauss 2001] for a pleasant exposition of these constructions.

Fact 2. *Conversely, any point in the Poincaré model obtained from initial data by using Euclidean ruler and compass can also be obtained using intrinsic* (*hyperbolic*) *ruler and compass.*

This is perhaps more surprising than Fact 1, since Euclidean manipulations in the Poincaré model can involve objects that are not actually in the hyperbolic plane, but rather on the boundary and the exterior of the disk. Fact 2 was apparently first proved in [Curtis 1990] — specifically in §6, but the whole article is recommended for its lucid discussion, references to earlier work, and a proof of the 90-year old result of D. Mordukhai-Boltovskoi that states exactly which lengths are constructible in hyperbolic geometry. (Warning: The "Klein conformal model" to which Curtis refers is a slight variation of the Poincaré model, rather than the projective Klein model explained on page 276.)

With this we can close our digression, confident that in talking about ruler-and-compass constructibility, it makes no difference whether we use intrinsic hyperbolic tools or Euclidean tools in the Poincaré model!

3. Correspondence between the hyperbolic and Euclidean Alhazen problems

We now prove the Theorem stated on page 274. We are given a disk D_H in the hyperbolic plane and a disk D_E in the Euclidean plane.

Lemma 1. *We can assume without loss of generality that D_H and D_E are centered at the origin O of the Cartesian plane.* (In the hyperbolic case, we understand the Cartesian plane as underlying the Poincaré model.)

This may seem obvious, but it merits discussion: the theorem does talk of arbitrary circles, after all. The proof of the lemma follows from three observations:

Any isometry T of either the Euclidean or the hyperbolic plane can be implemented with ruler and compass, in the following sense: Let T be defined by some known data, such as the images $T(a), T(b), T(c)$ of three noncollinear points a, b, c. Then, given any point x, one can construct $T(x)$ with ruler and compass, starting from x and the data defining T.

This is usually taken for granted for Euclidean constructions, and indeed it is easy to show — we leave it as an exercise. Your proof for the Euclidean case will quite likely carry over to the hyperbolic case (with intrinsic ruler and compass).

The second observation is that *any given disk is mapped by some isometry T to some disk centered at O*. This is because the hyperbolic and Euclidean planes are homogeneous: given two points in the Euclidean plane, there is an isometry taking one to the other — and of course such an isometry maps a disk to a disk. Similarly for the hyperbolic plane — having in mind hyperbolic isometries, of course.

The third observation ties it all together: The theorem asserts a constructibility-preserving bijection between D_H- and D_E-Alhazen configurations. If such a bijection is known for D_H and D_E of the special form in the lemma, it can be defined for *any* D_H and D_E, by using isometries to bridge between configurations in the old and the new D_H, and between configurations in the old and the new D_E. Because isometries are constructible, these bridges take constructible configurations to constructible configurations; and because they do so in both directions, they also take nonconstructible configurations to nonconstructible configurations.

This formally justifies the usual cavalier attitude about isometries when dealing with constructibility questions.

Let ϕ denote the map taking the Poincaré model to the Klein model, described in Figure 2 as the composition of stereographic projection and vertical projection. We need one more normalization.

Lemma 2. *We can assume, moreover, that D_H is taken to D_E under the Poincaré-to-Klein map ϕ.*

The radius of D_H cannot be tampered with,[2] but fortunately the radius of D_E can, by applying a homothety (scaling transformation). Homotheties preserve constructibility, being themselves constructible (same logic as in the third observation above). So we just scale D_E to make it coincide with $\phi(D_H)$.

Lemma 3. *The Poincaré-to-Klein correspondence ϕ, applied to Alhazen configurations in our normalized D_H and D_E, preserves constructibility.*

[2] The hyperbolic plane has no similarities other than isometries, so we cannot hope to reduce the hyperbolic Alhazen problem to a single circle size, as we're accustomed to doing with Euclidean problems. Another way to say this is that hyperbolic disks of different sizes are not scaled images of one another — in fact, the larger the circle, the greater the ratio between circumference and diameter!

Indeed, the action of ϕ on single points can be implemented with (Euclidean) ruler and compass. This is obvious for points on the red boundary circle,[3] since such points are just pulled halfway toward O — a homothety of ratio $\frac{1}{2}$. Now let P be a point inside the Poincaré disk. Use Poincaré (i.e., Euclidean) ruler and compass to draw any two hyperbolic lines crossing at P; this is possible by Fact 1. Mark the intersections of these lines with the Poincaré red circle (like points H_1 and H_2 in Figure 2). Apply ϕ to the four points thus determined on the Poincaré red circle, to find the corresponding points on the Klein red circle. Obtain $\phi(P)$ as the intersection of the two line segments connecting pairs of opposite points.

The key observation now is that *the hyperbolic Alhazen problem in D_H with initial data A, B has S as a solution if and only if the* ***Euclidean*** *Alhazen problem in D_E with initial data $\phi(A)$, $\phi(B)$ has $\phi(S)$ as a solution.* Here we're thinking of the solution as a single point — the reflection point on the circle between A and B.

The statement in italics applies to solutions in general, whether or not they are constructible. To wrap up the proof, we resort again to the bridge idea used for the first two lemmas. We spell it out here, since we didn't before. Because ϕ and its inverse are constructible, they preserve the constructibility status of solutions. That is, if we can get from A and B to the solution S, then we can get from $\phi(A)$ and $\phi(B)$ to $\phi(S)$, via A $(= \phi^{-1}(\phi(A)))$ and B and then S. Conversely, if we can get from $\phi(A)$ and $\phi(B)$ to $\phi(S)$, we can get from A and B to S. This finishes the proof of the Theorem.

Remark 1. What makes the Alhazen problem special is that we can write that boldfaced "***Euclidean***" above. For any problem, ϕ transforms hyperbolic solutions in the Poincaré model into hyperbolic solutions in the Klein model; but only here is the hyperbolic solution also a Euclidean solution, and only because we chose D_H and D_E judiciously in Lemmas 1 and 2, rendering irrelevant the difference between the Euclidean metric and the hyperbolic metric in the Klein model. That's why the proof fails for a problem such as finding a point T a third of the way from A to B (as already mentioned, the hyperbolic version of this problem is not constructible).

Remark 2. The bijection we have constructed between D_H- and D_E-Alhazen configurations is just a componentwise application of the Poincaré-to-Klein map ϕ, and is therefore very well behaved (homeomorphic, locally bi-Lipschitz). It follows that, for each circle radius, not only are there D_H-Alhazen configurations that are not solvable with ruler and compass, but in fact they are the rule. Solvable configurations are the exception — they form a set of measure zero in $D_H \times D_H$, corresponding to a set of measure zero of solvable configurations in the Euclidean case [Neumann 1998, p. 527]. This proves the Corollary.

[3] That these points are not in the hyperbolic plane itself doesn't matter: we're using them as stepping stones, and ϕ is obviously defined on them.

Acknowledgement

We thank the Mohler–Thompson Fund for sponsoring our research. The referee's suggestions transformed our mishmash of ideas into an article: a valuable contribution.

References

[Curtis 1990] R. R. Curtis, "Duplicating the cube and other notes on constructions in the hyperbolic plane", *J. Geom.* **39**:1-2 (1990), 38–59. MR 91i:51027 Zbl 0717.51018

[Elkin 1965] J. M. Elkin, "A deceptively easy problem", *Math. Teacher* **58** (1965), 194–199.

[Goodman-Strauss 2001] C. Goodman-Strauss, "Compass and straightedge in the Poincaré disk", *Amer. Math. Monthly* **108**:1 (2001), 38–49. MR 1857068 Zbl 1013.51010

[Martin 1975] G. E. Martin, *The foundations of geometry and the non-Euclidean plane*, Intext Educational, New York, 1975. Reprinted Springer, 1998. Zbl 0321.50001

[Neumann 1998] P. M. Neumann, "Reflections on reflection in a spherical mirror", *Amer. Math. Monthly* **105**:6 (1998), 523–528.

[Riede 1989] H. Riede, "Reflexion am Kugelspiegel, oder das Problem des Alhazen", *Praxis Math.* **31**:2 (1989), 65–70. MR 90g:51048

[Sabra 1982] A. I. Sabra, "Ibn al-Haytham's lemmas for solving 'Alhazen's Problem' ", *Arch. Hist. Exact Sci.* **26**:4 (1982), 299–324.

Received: 2011-01-18 Revised: 2011-06-24 Accepted: 2011-06-26

poirinat@aquinas.edu
Department of Mathematics, Aquinas College,
1607 Robinson Road SE, Grand Rapids 49506, United States

mcdanmic@aquinas.edu
Department of Mathematics, 1903 W. Michigan Ave.,
Western Michigan University, Kalamazoo, MI 49008-5248,
United States

INVOLVE 5:3 (2012)
dx.doi.org/10.2140/involve.2012.5.283

Bochner (p, Y)-operator frames

Mohammad Hasan Faroughi, Reza Ahmadi and Morteza Rahmani

(Communicated by David R. Larson)

Using the concepts of Bochner measurability and Bochner space, we introduce a continuous version of (p, Y)-operator frames for a Banach space. We also define independent Bochner (p, Y)-operator frames for a Banach space and discuss some properties of Bochner (p, Y)-operator frames.

1. Introduction and preliminaries

The concept of frames was first introduced in the context of nonharmonic Fourier series [Duffin and Schaeffer 1952], and after the publication of [Daubechies et al. 1986] it has found broad application in signal processing, image processing, data compression and sampling theory. In this paper we introduce *Bochner* (p, Y)*-operator frames*, which are the continuous version of (p, Y)-operator frames for a Banach space, introduced in [Cao et al. 2008]. The new frames also generalize the *continuous* p*-frames* introduced in [Faroughi and Osgooei 2011].

Throughout this paper H will be a Hilbert space and X will be a Banach space.

Definition 1.1. Let $\{f_i\}_{i\in I}$ be a sequence of elements of H. We say that $\{f_i\}_{i\in I}$ is a *frame* for H if there exist constants $0 < A \leq B < \infty$ such that for all $h \in H$

$$A\|h\|^2 \leq \sum_{i\in I} |\langle f_i, h\rangle|^2 \leq B\|h\|^2. \tag{1-1}$$

The constants A and B are called frame bounds. If A, B can be chosen so that $A = B$, we call this frame an A-tight frame and if $A = B = 1$ it is called a Parseval frame. If we only have the upper bound, we call $\{f_i\}_{i\in I}$ a Bessel sequence. If $\{f_i\}_{i\in I}$ is a Bessel sequence then the following operators are bounded:

$$T : l^2(I) \to H, \quad T(c_i) = \sum_{i\in I} c_i f_i, \tag{1-2}$$

$$T^* : H \to l^2(I), \quad T^*(f) = \{\langle f, f_i\rangle\}_{i\in I}, \tag{1-3}$$

MSC2010: 42C15, 46C05.

Keywords: Bochner (p, Y)-operator frame, Bochner measurability, Bochner space, Radon–Nikodym property.

called the *synthesis* and *analysis* operators, respectively. Hence the *frame operator* S, given by

$$Sf = TT^* f = \sum_{i \in I} \langle f, f_i \rangle f_i, \tag{1-4}$$

is also bounded.

The theory of frames has a continuous version, as follows.

Definition 1.2 [Rahimi et al. 2006]. Let (Ω, μ) be a measure space. Let $f : \Omega \to H$ be weakly measurable (i.e., for each $h \in H$, the mapping $\omega \to \langle f(\omega), h \rangle$ is measurable). Then f is called a *continuous frame* or *c-frame* for H if there exist constants $0 < A \le B < \infty$ such that for all $h \in H$

$$A\|h\|^2 \le \int_\Omega |\langle f(\omega), h \rangle|^2 d\mu \le B\|h\|^2. \tag{1-5}$$

In this context the synthesis operator $T_f : L^2(X, \mu) \to H$ is defined by

$$\langle T_f \phi, h \rangle = \int_X \phi(x) \langle f(x), h \rangle \, d\mu(x); \tag{1-6}$$

the analysis operator $T_f^* : H \to L^2(X, \mu)$ by

$$(T_f^* h)(x) = \langle h, f(x) \rangle, \quad x \in X; \tag{1-7}$$

and the frame operator by

$$S_f = T_f T_f^*. \tag{1-8}$$

By Theorem 2.5 in [Rahimi et al. 2006], S_f is positive, self-adjoint and invertible.

Suppose (Ω, Σ, μ) is a measure space, where μ is a positive measure.

Definition 1.3. A function $f : \Omega \to X$ is called *simple* if there exist $x_1, \dots, x_n \in X$ and $E_1, \dots, E_n \in \Sigma$ such that $f = \sum_{i=1}^n x_i \chi_{E_i}$, where $\chi_{E_i}(\omega) = 1$ if $\omega \in E_i$ and $\chi_{E_i}(\omega) = 0$ if $\omega \in E_i^c$. If $\mu(E_i)$ is finite whenever $x_i \neq 0$ then the simple function f is *integrable*, and the integral is then defined by

$$\int_\Omega f(\omega) \, d\mu(\omega) = \sum_{i=1}^n \mu(E_i) x_i.$$

Definition 1.4. A function $f : \Omega \to X$ is called *Bochner-measurable* if there exists a sequence of simple functions $\{f_n\}_{n=1}^\infty$ such that

$$\lim_{n \to \infty} \|f_n(\omega) - f(\omega)\| = 0, \quad \mu\text{-a.e.}$$

Definition 1.5. A Bochner-measurable function $f : \Omega \to X$ is called *Bochner-integrable* if there exists a sequence of integrable simple functions $\{f_n\}_{n=1}^\infty$ such that

$$\lim_{n \to \infty} \int_\Omega \|f_n(\omega) - f(\omega)\| \, d\mu(\omega) = 0.$$

In this case, $\int_E f(\omega)\, d\mu(\omega)$ is defined by

$$\int_E f(\omega)\, d\mu(\omega) = \lim_{n\to\infty} \int_E f_n(\omega)\, d\mu(\omega), \quad E \in \Sigma.$$

Definition 1.6. A Banach space X has the *Radon–Nikodym property* if, for every finite measure space (Ω, Σ, μ) and every (finitely additive) X-valued measure γ on (Ω, Σ) that has bounded variation and is absolutely continuous with respect to μ, there is a Bochner-integrable function $g : \Omega \to X$ such that

$$\gamma(E) = \int_E g(\omega)\, d\mu(\omega)$$

for every measurable set $E \in \Sigma$.

Remark 1.7. Suppose that (Ω, Σ, μ) is a measure space and X^* has the Radon–Nikodym property. Let $1 \le p \le \infty$. The *Bochner space* $L^p(\mu, X)$ is defined to be the Banach space of (equivalence classes of) X-valued Bochner-measurable functions F on Ω whose L^p norm is finite; here the L^p norm is defined by

$$\|F\|_p = \left(\int_\Omega \|F(\omega)\|^p\, d\mu(\omega)\right)^{1/p}$$

if p is finite, and by the essential supremum of $\|F(\omega)\|$ if $p = \infty$. In [Diestel and Uhl 1977; Cengiz 1998; Fleming and Jamison 2008, p. 51] it is proved that if q is such that $\frac{1}{p} + \frac{1}{q} = 1$, then $L^q(\mu, X^*)$ is isometrically isomorphic to $(L^p(\mu, X))^*$ if and only if X^* has the Radon–Nikodym property. This isometric isomorphism

$$\psi : L^q(\mu, X^*) \to (L^p(\mu, X))^*$$

takes $g \in L^q(\mu, X^*)$ to ϕ_g, the linear map defined by

$$\phi_g(f) = \int_\Omega g(\omega)(f(\omega))\, d\mu(\omega), \quad f \in L^p(\mu, X).$$

So for all $f \in L^p(\mu, X)$ and $g \in L^q(\mu, X^*)$ we have

$$\psi(g)(f) = \langle f, \psi(g)\rangle = \int_\Omega g(\omega)(f(\omega))\, d\mu(\omega) = \int_\Omega \langle f(\omega), g(\omega)\rangle\, d\mu(\omega).$$

In the following, we use the notation $\langle f, g\rangle$ instead of $\langle f, \psi(g)\rangle$, so for all $f \in L^p(\mu, X)$ and $g \in L^q(\mu, X^*)$

$$\langle f, g\rangle = \int_\Omega \langle f(\omega), g(\omega)\rangle\, d\mu(\omega).$$

Hilbert spaces have the Radon–Nikodym property, so in particular, if H is a Hilbert space then $(L^p(\mu, H))^*$ is isometrically isomorphic to $L^q(\mu, H)$. So, for all $f \in L^p(\mu, H)$ and $g \in L^q(\mu, H)$, we have

$$\langle f, g\rangle = \int_\Omega \langle f(\omega), g(\omega)\rangle \, d\mu(\omega),$$

in which $\langle f(\omega), g(\omega)\rangle$ does not mean the inner product of elements $f(\omega)$, $g(\omega)$ in H, but

$$\langle f(\omega), g(\omega)\rangle = \nu(g(\omega))(f(\omega)),$$

where $\nu : H \to H^*$ is the isometric isomorphism between H and H^*.

Lemma 1.8. *Let (Ω, Σ, μ) be a measure space and suppose there exists $k > 0$ such that $\mu(E) \geq k$ for every nonempty measurable set E of Ω. For every $\omega \in \Omega$, define $P_\omega : L^p(\mu, X) \to X$, $P_\omega(G) = G(\omega)$. Then $\|P_\omega\| \leq k^{-1/p}$.*

Proof. For a fix $\omega_0 \in \Omega$, put

$$\Delta = \{\omega \in \Omega \mid \|G(\omega)\| \geq \|G(\omega_0)\|\}.$$

Then

$$\|G\|_p^p = \int_\Omega \|G(\omega)\|^p \, d\mu(\omega) \geq \int_\Delta \|G(\omega)\|^p \, d\mu(\omega) \geq \mu(\Delta)\|G(\omega_0)\|^p \geq k\|G(\omega_0)\|^p.$$

Hence

$$\|P_{\omega_0}\| = \sup_{\|G\|_p \leq 1} \|P_{\omega_0}(G)\| = \sup_{\|G\|_p \leq 1} \|G(\omega_0)\| \leq \sup_{\|G\|_p \leq 1} k^{-1/p}\|G\|_p = k^{-1/p}. \quad \square$$

2. Bochner (p, Y)-Bessel mappings for X

Throughout this section and the next we will work with a second Banach space Y in addition to X. We denote by $B(X, Y)$ the space of bounded operators from X to Y.

Definition 2.1. Let $1 < p < \infty$, and let $F : \Omega \to B(X, Y)$ be a map; we write F_ω for $F(\omega)$. We say that F is a *Bochner (p, Y)-Bessel mapping for X* if the following conditions are met:

(i) For each $x \in X$, the mapping $\omega \mapsto F_\omega(x)$ from Ω into Y is Bochner-measurable.

(ii) There exists a positive constant B such that

$$\|F_.(x)\|_p \leq B\|x\| \quad \text{for all } x \in X, \tag{2-1}$$

where

$$\|F_.(x)\|_p = \left(\int_\Omega \|F_\omega(x)\|^p \, d\mu\right)^{1/p}. \tag{2-2}$$

We denote by $B_X^p(Y)$ the set of all Bochner (p, Y)-Bessel mappings for X. It is easy to see that this set is closed under addition (defined in the obvious way: for $F, K \in B_X^p(Y)$, the sum $F+K$ satisfies $(F+K)_\omega(x) = F_\omega(x) + K_\omega(x)$ for all $x \in X$ and $\omega \in \Omega$) and under multiplication by scalars. Thus $B_X^p(Y)$ is a vector space. We give it a norm as follows. The *Bessel bound of* $F \in B_X^p(Y)$ is the number

$$B_F = \inf\{B > 0 : B \text{ satisfies (2-1)}\}.$$

For every $F \in B_X^p(Y)$, define $R_F : X \to L^p(\mu, Y)$ by $x \mapsto F_\cdot(x)$. This is clearly a linear map; we should that it is also bounded. For every $F \in B_X^p(Y)$,

$$\|R_F(x)\|_p = \|F_\cdot(x)\|_p \leq B\|x\|, \tag{2-3}$$

for any B satisfying (2-1). Together with the linearity of R_F this implies that

$$\|R_F\| \leq B_F; \tag{2-4}$$

that is, $R_F \in B(X, L^p(\mu, Y))$. Now set

$$\|F\|_p = \|R_F\|. \tag{2-5}$$

By (2-4), $\|F\|_p \leq B_F$. It is easy to show that this gives a norm on $B_X^p(Y)$.

Theorem 2.2. *Let* (Ω, Σ, μ) *be a measure space and suppose there exists* $k > 0$ *such that* $\mu(E) \geq k$ *for every nonempty measurable set* E *of* Ω. *For every* $1 < p < \infty$, *the mapping*

$$\Lambda : B_X^p(Y) \to B(X, L^p(\mu, Y))$$

given by $\Lambda(F) = R_F$ *is a linear isometric isomorphism, and* $B_X^p(Y)$ *is a Banach space over* $\mathbb{C}$.

Proof. Clearly, the mapping Λ is a linear isometry from $B_X^p(Y)$ into $B(X, L^p(\mu, Y))$. Next we prove that Λ is surjective.

Choose $\omega \in \Omega$. For every $A \in B(X, L^p(\mu, Y))$, define $F_\omega^A : X \to Y$ by

$$F_\omega^A(x) = P_\omega(A(x)) = A(x)(\omega), \quad x \in X.$$

By Lemma 1.8, we have $\|P_\omega\| \leq k^{-1/p}$; hence $F_\omega^A \in B(X, Y)$ for all $\omega \in \Omega$. Now, consider the mapping

$$F^A : \Omega \to B(X, Y)$$

given by $\omega \mapsto F_\omega^A$. Since $F_\cdot^A(x) = A(x)(\,\cdot\,) : \Omega \to Y$ for each $x \in X$, the mapping $\omega \mapsto F_\omega^A(x)$ from Ω into Y is Bochner-measurable and

$$\|A(x)\|_p = \int_\Omega \|A(x)(\omega)\|^p \, d\mu(\omega) = \int_\Omega \|F_\omega^A(x)\|^p \, d\mu(\omega) = \|F_\cdot^A(x)\|_p.$$

Therefore

$$\|F_\cdot^A(x)\|_p = \|A(x)\|_p \leq \|A\|\,\|x\|.$$

Hence $F^A \in B^p_X(Y)$. Also, for all $\omega \in \Omega$ we have $R_{F^A}(x)(\omega) = F^A_\omega(x) = A(x)(\omega)$. Thus $R_{F^A}(x) = A(x)$ for all $x \in X$. This shows that $\Lambda(F^A) = R_{F^A} = A$; thus Λ is surjective and so bijective. Consequently, $B^p_X(Y)$ is isometrically isomorphic to the Banach space $B(X, L^p(\mu, Y))$. Therefore, $B^p_X(Y)$ is a Banach space over $\mathbb{C}$. □

Theorem 2.3. *Let $1 < p < \infty$ and $F \in B^p_X(Y)$. Then, for every $y^* \in Y^*$, the mapping $F^*_{\cdot}(y^*) : \Omega \to X^*$, $F^*_{\cdot}(y^*)(\omega) = F^*_\omega(y^*)$ is a Bochner pg-Bessel mapping for X with respect to $\mathbb{C}$.*

Proof. Let $y^* \in Y^*$ and $x \in X$. Clearly for each $x \in X$ the map $\omega \mapsto \langle x, F^*_\omega(y^*)\rangle$ from Ω into $\mathbb{C}$ is measurable and

$$\begin{aligned}\int_\Omega |\langle x, F^*_\omega(y^*)\rangle|^p \, d\mu(\omega) &= \int_\Omega |\langle F_\omega(x), y^*\rangle|^p \, d\mu(\omega)\\ &\le (\|y^*\|^p)\left(\int_\Omega \|F_\omega(x)\|^p \, d\mu(\omega)\right)\\ &\le \|y^*\|^p B^p_F \|x\|^p. \end{aligned}$$

□

Theorem 2.4. *Let (Ω, μ) be a σ-finite measure space with positive measure μ and let $\Omega = \bigcup_{n\in\mathbb{N}} K_n$ with $K_n \subseteq K_{n+1}$. Let $1 < p < \infty$, $\frac{1}{p} + \frac{1}{q} = 1$ and $F : \Omega \to B(X, Y)$. The following assertions are equivalent*:

(i) *$F \in B^p_X(Y)$.*

(ii) *For each $x \in X$, $\int_\Omega \|F_\omega(x)\|^p \, d\mu(\omega) < \infty$.*

(iii) *For each $G \in L^q(Y^*)$, $\sup_{\|x\|\le 1} |\int_\Omega \langle x, F^*_\omega(G(\omega))\rangle \, d\mu(\omega)| < \infty$.*

(iv) *The operator $S_F : L^q(Y^*) \to X^*$ defined by*

$$\langle x, S_F(G)\rangle = \int_\Omega \langle x, F^*_\omega(G(\omega))\rangle \, d\mu(\omega) \quad \textit{for } x \in X$$

is well defined and bounded.

Proof. (i) ⇒ (ii) This is obvious.

(ii) ⇒ (i) Define $A_n : X \to L^p(Y)$ by $A_n(x)(\omega) = \chi_{K_n}(\omega) F_\omega(x)$. For every $n \in \mathbb{N}$, we have

$$\|A_n\| = \sup_{\|x\|\le 1} \|A_n(x)\|_p \le \|F_\omega\|.$$

Hence, for all $n \in \mathbb{N}$, $A_n \in B(X, L^p(Y))$. By the definition of R_F, for every $n \in \mathbb{N}$,

$$\begin{aligned}\|(R_F - A_n)(x)\|^p_p &= \int_\Omega \|R_F(x)(\omega) - A_n(x)(\omega)\|^p \, d\mu(\omega)\\ &= \int_\Omega \|F_\omega(x) - \chi_{K_n}(\omega) F_\omega(x)\|^p \, d\mu(\omega)\\ &= \int_{\Omega - K_n} \|F_\omega(x)\|^p \, d\mu(\omega).\end{aligned}$$

This converges to 0 as $n \to \infty$, proving that $\lim_{n\to\infty} A_n(x) = R_F(x)$ for all $x \in X$. By the Banach–Steinhaus theorem, $R_F \in B(X, L^p(Y))$ and $\|R_F\| = \sup \|A_n\| < \infty$. Hence $F \in B_X^p(Y)$.

(i) ⇒ (iii) Let $G \in L^q(\mu, Y^*)$ be arbitrary. By the Hölder inequality, we have

$$\begin{aligned}
&\sup_{\|x\|\le 1}\left|\int_\Omega \langle x, F_\omega^*(G(\omega))\rangle \, d\mu(\omega)\right| \\
&\qquad = \sup_{\|x\|\le 1}\left|\int_\Omega \langle F_\omega(x), G(\omega)\rangle \, d\mu(\omega)\right| \\
&\qquad \le \sup_{\|x\|\le 1}\left(\int_\Omega \|F_\omega(x)\|^p \, d\mu(\omega)\right)^{1/p}\left(\int_\Omega \|G\omega\|^q \, d\mu(\omega)\right)^{1/q} \le B_F\|G\|_q < \infty.
\end{aligned}$$

(iii) ⇒ (iv) Clearly S_F is well defined and by the proof of (i) ⇒ (iii) we have

$$\|S_F\| = \sup_{\|G\|_q\le 1} \|S_F(G)\| = \sup_{\|G\|_q\le 1}\ \sup_{\|x\|\le 1} \langle S_F(G), x\rangle \le B_F < \infty.$$

(iv) ⇒ (i) Take $G \in L^q(\mu, Y^*)$ such that $\|G(\omega)\| = 1$ for every $\omega \in \Omega$ and

$$\|F_\omega(x)\| = \langle F_\omega(x), G(\omega)\rangle = \langle x, F_\omega^*(G(\omega))\rangle \quad \text{for all } x \in X.$$

Define $\alpha_n : \Omega \to Y^*$ by $\alpha_n(\omega) = \chi_{K_n}(\omega)\|F_\omega(x)\|^{p-1}G(\omega)$. Then

$$\begin{aligned}
\|\alpha_n\|_q &= \left(\int_\Omega \|\chi_{K_n}(\omega)\|F_\omega(x)\|^{p-1}G(\omega)\|^q \, d\mu(\omega)\right)^{1/q} \\
&= \left(\int_{K_n} \|F_\omega(x)\|^{q(p-1)} \, d\mu(\omega)\right)^{1/q} = \left(\int_{K_n} \|F_\omega(x)\|^p \, d\mu(\omega)\right)^{1/q}.
\end{aligned}$$

Now, we have

$$\begin{aligned}
\int_{K_n} \|F_\omega(x)\|^p \, d\mu(\omega) &= \int_{K_n} \langle x, \|F_\omega(x)\|^{p-1} F_\omega^*(G(\omega))\rangle \, d\mu(\omega) \\
&= \int_\Omega \big\langle x, \chi_{K_n}(\omega)\|F_\omega(x)\|^{p-1} F_\omega^*(G(\omega))\big\rangle \, d\mu(\omega) = \langle x, S_F(\alpha_n)\rangle \\
&\le \|x\|\,\|S_F\|\,\|\alpha_n\|_q = \|x\|\,\|S_F\|\left(\int_{K_n} \|F_\omega(x)\|^p \, d\mu(\omega)\right)^{1/q}.
\end{aligned}$$

Thus

$$\left(\int_{K_n} \|F_\omega(x)\|^p \, d\mu(\omega)\right)^{1/p} \le \|x\|\,\|S_F\|. \tag{2-6}$$

By letting $n \to \infty$ in (2-6), we get $F \in B_X^p(Y)$. □

3. Bochner (p, Y)-operator frames

Definition 3.1. Let $1 < p < \infty$. A mapping $F : \Omega \to B(X, Y)$ is called a *Bochner (p, Y)-operator frame* for X if the following conditions hold:

(i) For each $x \in X$, the mapping $\omega \mapsto F_\omega(x)$ from Ω into Y is Bochner-measurable.

(ii) There exist positive constants A and B such that

$$A\|x\| \le \|F_{.}(x)\|_p \le B\|x\| \quad \text{for all } x \in X, \tag{3-1}$$

where $\|F_{.}(x)\|_p$ is as in (2-2). The *lower* and *upper bounds* of F are then given by

$$A_F = \sup\{A > 0 : A \text{ satisfies (3-1)}\}, \quad B_F = \inf\{B > 0 : B \text{ satisfies (3-1)}\},$$

We denote by $F_X^p(Y)$ the set of all Bochner (p, Y)-operator frames for X.

Definition 3.2. A Bochner (p, Y)-operator frame F is called *tight* if $A_F = B_F$. If $A_F = B_F = 1$, we call F *normalized*. We denote by $TF_X^p(Y)$ and $NF_X^p(Y)$, respectively, the sets of all tight and normalized Bochner (p, Y)-operator frames for X.

Corollary 3.3. *Let $F \in B_X^p(Y)$.*

(i) *$F \in F_X^p(Y)$ if and only if R_F is bounded below if and only if R_F^* is surjective.*

(ii) *$F \in TF_X^p(Y)$ if and only if R_F is a scaled isometry.*

Lemma 3.4. (i) *If $F \in B_X^p(Y)$ then $R_F^*\psi = S_F$.*

(ii) *If Y is reflexive then $L^p(\mu, Y)$ is reflexive.*

Proof. (i) For all $g \in L^q(\mu, Y^*)$ and $x \in X$, we have

$$\begin{aligned}\langle x, R_F^*\psi(g)\rangle = \langle R_F x, \psi(g)\rangle &= \int_\Omega \langle F_\omega(x), g(\omega)\rangle \, d\mu(\omega)\\ &= \int_\Omega \langle x, F_\omega^*(g(\omega))\rangle \, d\mu(\omega) = \langle x, S_F g\rangle.\end{aligned}$$

(ii) Let $J_Y : Y \to Y^{**}$ be the canonical mapping. Suppose that Y is reflexive, that is $J_Y(Y) = Y^{**}$. For every $f \in L^p(\mu, Y)$, define $L^p(J_Y)(f(\omega)) = J_Y f(\omega)$, $\omega \in \Omega$. This gives a bijection $L^p(J_Y) : L^p(\mu, Y) \to L^p(\mu, Y^{**})$. By using Remark 1.7, we know that the mapping $\psi : L^q(\mu, Y^*) \to (L^p(\mu, Y))^*$ is a bijective bounded operator and so the adjoint $\psi^* : (L^p(\mu, Y))^{**} \to (L^q(\mu, Y^*))^*$ is bijective.
By using Remark 1.7 again, we obtain a bijective bounded operator

$$\psi' : L^p(\mu, Y^{**}) \to (L^q(\mu, Y^*))^*$$

such that for all $f \in L^p(\mu, Y^{**})$ and $g \in L^q(\mu, Y^*)$

$$\langle f, \psi' g\rangle = \int_\Omega \langle f(\omega), g(\omega)\rangle \, d\mu(\omega).$$

For all $f \in L^p(\mu, Y)$, $g \in L^q(\mu, Y^*)$ we have

$$\big\langle g, (\psi^* \circ J_{L^p(\mu,Y)}) f\big\rangle = \langle \psi(g), J_{L^p(\mu,Y)} f\rangle = \langle f, \psi(g)\rangle = \int_\Omega \langle f(\omega), g(\omega)\rangle \, d\mu(\omega)$$

and

$$\begin{aligned}\big\langle g, (\psi' \circ L^p(J_Y)) f\big\rangle &= \langle g, (\psi'(J_Y f(\cdot)))\rangle \\ &= \int_\Omega \langle g(\omega), J_Y f(\omega)\rangle \, d\mu(\omega) \\ &= \int_\Omega \langle f(\omega), g(\omega)\rangle \, d\mu(\omega).\end{aligned}$$

Therefore, $\psi^* \circ J_{L^p(\mu,Y)} = \psi' \circ L^p(J_Y)$ and hence $J_{L^p(\mu,Y)} = (\psi^*)^{-1} \circ \psi' \circ L^p(J_Y)$, which is a bijection. Hence $L^p(\mu, Y)$ is reflexive. □

Theorem 3.5. *Let $F \in B_X^p(Y)$, $G \in F_X^p(Y)$ and $\|F\|_p \le A_G$. Then*

$$F \pm G \in F_X^p(Y).$$

Proof. For each $x \in X$, we have

$$\|(F \pm G)_.(x)\|_p = \|F_.(x) \pm G_.(x)\|_p \ge \|G_.(x)\|_p - \|F_.(x)\|_p \ge (A_G - \|F\|_p)\|x\|$$

and

$$\|(F \pm G)_.(x)\|_p \le (\|F\|_p + \|G\|_p)\|x\|.$$

So $F \pm G \in F_X^p(Y)$. □

Theorem 3.6. *Let $F \in F_X^p(Y)$. Then for each $x^* \in X^*$, there exists an element $G \in L^p(\mu, Y^*)$ such that*

$$\langle y, x^*\rangle = \int_\Omega \langle y, F_\omega^*(G(\omega))\rangle \, d\mu(\omega), \qquad y \in X.$$

Proof. By Lemma 3.4, we have $R_F^* \psi = S_F$. Since $F \in F_X^p(Y)$, it follows from Corollary 3.3 that R_F^* is surjective. Thus the operator $S_F : L^q(\mu, Y^*) \to X^*$ is a surjection. Let $x^* \in X^*$; then there exists a $G \in L^p(\mu, Y^*)$ such that $x^* = S_F(G)$, so

$$\langle y, x^*\rangle = \int_\Omega \langle y, F_\omega^*(G(\omega))\rangle \, d\mu(\omega), \qquad y \in X. \qquad \square$$

Definition 3.7. A Bochner (p, Y)-operator frame for X is called *independent* if the operator S_F is injective, i.e., if for every $f \neq 0$ there exists $x \in X$ such that

$$\int_\Omega \langle x, F_\omega^*(f(\omega))\rangle \, d\mu(\omega) \neq 0.$$

We denote by $IF_X^p(Y)$ the set of all independent Bochner (p, Y)-operator frames for X.

Theorem 3.8. *Let F be an independent Bochner (p, Y)-operator frame for X. Then R_F is invertible.*

Proof. We already know that S_F is injective. By Lemma 3.4 and Corollary 3.3, we know that R_F^* is bijective. Hence R_F is invertible. □

Theorem 3.9. *Let (Ω, Σ, μ) be a measure space and suppose there exists $k > 0$ such that $\mu(E) \geq k$ for every nonempty measurable set E of Ω. For each $F \in IF_X^p(Y)$, there exists a unique Bochner (q, Y^*)-operator frame Q for X^* such that for all $y \in X$*

$$\langle y, x^*\rangle = \int_\Omega \langle y, F_\omega^* R_Q x^*(\omega)\rangle \, d\mu(\omega).$$

Proof. Let F be an independent Bochner (p, Y)-operator frame for X. Then Theorem 3.8 yields that the operator R_F is invertible, so by Lemma 3.4, S_F is invertible. We can define $Q_\omega = P_\omega S_F^{-1}, \omega \in \Omega$, where $P_\omega : L^q(\mu, Y^*) \to Y^*$ is defined by $P_\omega(G) = G(\omega)$. By Lemma 1.8, P_ω is bounded. Therefore $Q_\omega \in B(X^*, Y^*)$, $\omega \in \Omega$. For each $x^* \in X^*$, we have $Q_.(x^*) = S_F^{-1}(x^*)$, so for each $x^* \in X^*$, the mapping $\omega \mapsto Q_\omega(x^*)$ is Bochner-measurable and

$$\frac{1}{\|S_F\|}\|x^*\| \leq \left(\int_\Omega \|Q_\omega(x^*)\|^q \, d\mu\right)^{1/q} = \|S_F^{-1}(x^*)\| \leq \|S_F^{-1}\|\|x^*\|.$$

Hence, Q is a Bochner (q, Y^*)-operator frame for X^* with bounds $\|S_F\|^{-1}$ and $\|S_F^{-1}\|$. By the definition of Q, we obtain that $R_Q = S_F^{-1}$ and so $x^* = S_F R_Q x^*$, $x^* \in X^*$. Thus

$$\langle y, x^*\rangle = \int_\Omega \langle y, F_\omega^* R_Q x^*(\omega)\rangle \, d\mu(\omega), \quad y \in X.$$

Next, we will show the uniqueness of Q. Let W be a Bochner (q, Y^*)-operator frame for X^* such that for all $y \in X$

$$\langle y, x^*\rangle = \int_\Omega \langle y, F_\omega^* R_W x^*(\omega)\rangle \, d\mu(\omega), \quad x^* \in X^*.$$

Thus $S_F R_W = I_{X^*}$, or $R_W = S_F^{-1} = R_Q$. Therefore, $W = Q$. □

References

[Cao et al. 2008] H.-X. Cao, L. Li, Q.-J. Chen, and G.-X. Ji, "(p, Y)-operator frames for a Banach space", *J. Math. Anal. Appl.* **347**:2 (2008), 583–591. MR 2009h:46024 Zbl 05344335

[Cengiz 1998] B. Cengiz, "The dual of the Bochner space $L^p(\mu, E)$ for arbitrary μ", *Turkish J. Math.* **22**:3 (1998), 343–348. MR 99k:46061 Zbl 0930.46034

[Daubechies et al. 1986] I. Daubechies, A. Grossmann, and Y. Meyer, "Painless nonorthogonal expansions", *J. Math. Phys.* **27**:5 (1986), 1271–1283. MR 87e:81089 Zbl 0608.46014

[Diestel and Uhl 1977] J. Diestel and J. J. Uhl, Jr., *Vector measures*, Mathematical Surveys **15**, American Mathematical Society, Providence, RI, 1977. MR 56 #12216 Zbl 0369.46039

[Duffin and Schaeffer 1952] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series", *Trans. Amer. Math. Soc.* **72** (1952), 341–366. MR 13,839a Zbl 0049.32401

[Faroughi and Osgooei 2011] M. H. Faroughi and E. Osgooei, "Continuous p-Bessel mappings and continuous p-frames in Banach spaces", *Involve* **4**:2 (2011), 167–186. MR 2012m:42045 Zbl 1235.42026

[Fleming and Jamison 2008] R. J. Fleming and J. E. Jamison, *Isometries on Banach spaces: vector-valued function spaces*, vol. 2, Monographs and Surveys in Pure and Applied Mathematics **138**, CRC, Boca Raton, FL, 2008. MR 2009i:46001 Zbl 1139.46001

[Rahimi et al. 2006] A. Rahimi, A. Najati, and Y. N. Dehghan, "Continuous frames in Hilbert spaces", *Methods Funct. Anal. Topology* **12**:2 (2006), 170–182. MR 2007d:42061 Zbl 1120.42019

Received: 2011-04-16 Accepted: 2012-01-21

mhfaroughi@yahoo.com
Department of Pure Mathematics,
Faculty of Mathematical Science, University of Tabriz,
29 Bahman Street, Tabriz 5166614766, Iran

reza.ahmadi84@yahoo.com
Department of Pure Mathematics,
Faculty of Mathematical Science, University of Tabriz,
29 Bahman Street, Tabriz 5166614766, Iran

m_rahmani26@yahoo.com
Department of Pure Mathematics,
Faculty of Mathematical Science, University of Tabriz,
29 Bahman Street, Tabriz 5166614766, Iran

Department of Mathematics, Ilam University,
P.O. Box 69315516, Ilam, Iran

INVOLVE 5:3 (2012)
dx.doi.org/10.2140/involve.2012.5.295

k-furcus semigroups

Nicholas R. Baeth and Kaitlyn Cassity

(Communicated by Scott Chapman)

A bifurcus semigroup is a semigroup in which every nonunit nonatom can be written as the product of exactly two atoms. We generalize this notion to k-furcus semigroups: every element that can be factored as the product of at least k nonunits can be factored as the product of exactly k atoms. We compute some factorization-theoretic invariants of k-furcus semigroups that generalize the bifurcus results. We then define two variations on the k-furcus property, one stronger (presumabaly strictly) and the other strictly weaker than the k-furcus property.

1. Introduction

Vadim Ponomarenko and coworkers [Adams et al. 2009] introduced and studied the notion of *bifurcus semigroups*, a class of semigroups with "bad" factorization properties: a semigroup S is bifurcus if every nonunit nonatom can be bifurcated, that is, expressed as the product of exactly two atoms in S. They gave examples, showed that certain important families of semigroups cannot be bifurcus, and calculated several factorization-theoretic invariants of bifurcus semigroups. Further examples of bifurcus semigroups can be found in [Baeth et al. 2011]. Our goal is to generalize the concept of bifurcus and to provide analogous results for what we call *k-furcus semigroups*. We also give in Section 3 two modified definitions, one which appears to strengthen and one which weakens the original definition. Finally, in Section 4, we consider irreducible divisor graphs — a graphical interpretation of the factorization of an element in a semigroup — of elements in k-furcus semigroups.

First, some basic background. The reader is referred to [Geroldinger and Halter-Koch 2006] for a thorough treatment of factorization theory.

A *semigroup* is a set together with an associative operation. A nonunit a of a semigroup S is an *atom* if it is impossible to write $a = b \cdot c$ with b and c nonunits. The set $\mathcal{A}(S)$ denotes the set of all atoms of S. We will restrict our attention to *atomic* semigroups, those in which every element can be expressed as a (finite)

MSC2010: 11Y05.
Keywords: semigroups, factorization.

product of atoms. We now define several important invariants which describe how unique or nonunique factorization is in a given semigroup.

An element $a \in S$ is a *strong atom* if whenever $a^m = bc$ for some $b, c \in S$ with $b \neq 1$, then $b = \epsilon a^n$ for some unit ϵ and some integer $n \leq m$. If $x \in S$, then $\mathfrak{L}(x) = \{n : x = a_1 a_2 \cdots a_n$ with each a_i an atom of $S\}$ is called the *set of factorization lengths* of x. The *elasticity* of an element x, defined by

$$\rho(x) = \frac{\sup \mathfrak{L}(x)}{\inf \mathfrak{L}(x)},$$

gives a coarse measure of how far away x is from having unique factorization. The *elasticity* of the semigroup S is then $\rho(S) = \sup\{\rho(x) : x \text{ is a nonunit of S}\}$. If $\mathfrak{L}(x) = \{t_1, t_2, \dots\}$ is the set of factorization lengths of x with $t_i < t_{i+1}$ for each i, the *delta set* of x is defined to be $\Delta(x) = \{t_{i+1} - t_i : t_i, t_{i+1} \in \mathfrak{L}(x)\}$ and $\Delta(S) = \bigcup \Delta(x)$. If $\Lambda = \{\min \mathfrak{L}(x) : x$ is a nonunit of $S\}$, then the *critical length* of S is $cr(S) = \max \Lambda + 1$. The *catenary degree* of S, denoted $C(S)$, is the smallest integer N such that for all $a \in S$, and for any two factorizations z and z' of a, there exists factorizations $z_0, \dots, z_k$ of a such that for all $i \in [1, k]$, z_i arises from z_{i-1} by replacing at most N atoms from z_{i-1} by at most N new atoms; that is, $d(z_i, z_{i-1}) \leq N$. Finally, we define the *tame degree* $t(S)$ of the semigroup S to be the smallest natural number N such that whenever $a \in S$ and x is an atom of S occurring in some factorization of a, given a factorization z of a, there exists a factorization z' of a containing x such that $d(z, z') \leq N$.

2. k-furcus semigroups

Let S be an atomic semigroup and let $k \geq 2$ be an integer. We say S is *k-furcus* if whenever an element can be factored as a product of at least k nonunits, then it can be factored as the product of exactly k atoms of S. Note that when $k = 2$, a semigroup is k-furcus if and only if it is bifurcus.

The following result generalizes (2)–(9) of [Adams et al. 2009, Theorem 1.1] for k-furcus semigroups, and can be proved by straightforward modifications of the arguments in that paper.

Theorem 2.1. *Let S be a nontrivial k-furcus semigroup, and let $0 \neq x \in S$ be a nonunit nonatom. Then*:

(1) *S contains no strong atoms.*

(2) *If $k \geq \sup \mathfrak{L}(x)$, then* $[k, \sup \mathfrak{L}(x)] \subseteq \mathfrak{L}(x) \subseteq [2, \sup \mathfrak{L}(x)]$.

(3) $k\rho(x) \in \mathbb{N} \cup \{\infty\}$.

(4) $\rho(S) = \infty$.

(5) (a) *If* $k \in \{2, 3\}$, *then* $\Delta(S) = \{1\}$.
(b) *If* $k > 3$, *then* $\{1\} \subseteq \Delta(S) \subseteq \{1, 2, \dots, k-2\}$.

(6) $3 \le C(S) \le k+1$.

(7) $t(S) = \infty$.

(8) $3 \le cr(S) \le k+1$.

We note that the statements (2), (5), and (8) of Theorem 2.1 are strictly weaker than their analogs (3), (6), and (9) from [Adams et al. 2009, Theorem 1.1]. Any improvements on these statements would require knowledge of how elements with no factorizations of length k or greater can be written as products of atoms.

Suppose that S is a k-furcus semigroup and that m is an integer larger than k. Further suppose that $x \in S$ can be written as the product of at least m nonunits of S. Then $k \le m \le \sup \mathfrak{L}(x)$ and thus, by Theorem 2.1(2), $m \in \mathfrak{L}(x)$ and so x can be factored into exactly m atoms. Therefore:

Corollary 2.2. *If a semigroup S is k-furcus then S is m-furcus for every $m \ge k$.*

Example 2.3. As shown in [Adams et al. 2009, Section 2], the following semigroups are bifurcus, and hence (by Corollary 2.2) k-furcus for all $k \ge 2$:

(1) $n\mathbb{Z}$, where n is not a prime power;

(2) $m\mathbb{Z} \times n\mathbb{Z}$ for natural numbers $m, n > 1$;

(3) the multiplicative subsemigroup of $n \times n$ matrices with all entries identical integers, for n not a prime power.

To be more concrete, consider the semigroup $n\mathbb{Z}$, where $n = pqr$ with p and q distinct primes, and suppose that $x \in n\mathbb{Z}$ can be written as $x = (nx_1)(nx_2)\cdots(nx_k)$ for some $k \ge 2$. Then we can factor x in $\mathbb{Z}$ as $x = p^a q^b r^k s$ where $a, b \ge k$ and $p, q \nmid s$. Then we can factor x as $x = (np^{a-k}s)(nq^{b-k})(n)^{k-2}$ as a product of exactly k atoms of $n\mathbb{Z}$. Therefore, $n\mathbb{Z}$ is k-furcus for all $k \ge 2$.

We now provide an example of a k-furcus semigroup that is not m-furcus for any $m < k$, showing that the term k-furcus properly extends the term bifurcus. We thank Vadim Ponomarenko (private communication) for providing this example.

Example 2.4. Consider k distinct primes $p_1, p_2, \dots, p_k$ and let $N = p_1 p_2 \cdots p_k$. Define S to be the multiplicative semigroup with the infinitely many generators $Np_1^{a_1}, Np_2^{a_2} \dots, Np_k^{a_k}$, where each a_i is a nonnegative integer. If x is an element of S that can be written as the product of at least k elements of S, then $x = p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k}$ for some $a_1, a_2, \dots, a_k$. Note that $a_i \ge k$ for each i because $N = p_1 p_2 \cdots p_k$ divides (in $\mathbb{Z}$) every element in S. Now we can write $x = (Np_1^{(a_1-k)})(Np_2^{(a_2-k)})\cdots(Np_k^{(a_k-k)})$ as a product of exactly k atoms and hence S is k-furcus. However, we shall see that it is impossible to write x as the product of less than k atoms.

Suppose $x = b_1 b_2 \cdots b_{k-1}$, where each b_i is an atom of S. Since p_1 occurs at least k times in the factorization of x, it must occur at least twice in a factorization of some b_i and hence each factorization of x must contain, for each i, an atom of the form $Np_i^{c_i}$ with $c_i \geq 1$. Thus every factorization of x must have length at least k. Since S is k-furcus we know that N will occur at least k times in any factorization of x. By factoring x into $k-1$ elements we can see that N will still have to occur k times, so x can not be written as a product of atoms of length less than k.

In [Adams et al. 2009], it is shown that several important families of semigroups are not bifurcus. Straightforward modifictions show that these same families of semigroups cannot be k-furcus for any $k \geq 2$.

Proposition 2.5. *These classes of semigroups are not k-furcus for any $k \geq 2$:*

(1) *block monoids $B(G_0)$ for any subset G_0 of an abelian group G*;

(2) *Krull monoids*;

(3) *diophantine monoids.*

3. Variations of k-furcus semigroups

In this section we consider variations of k-furcus semigroups, one weaker and one stronger.

We call a semigroup S *quasi k-furcus* if every nonunit has a factorization of length at most k. This definition is motivated by the following example, which follows from [Banister et al. 2007, Theorem 2.3]:

Example 3.1. If $M(a, b) = \{a + kb : k \in \mathbb{N}_0\} \cup \{1\}$ is an arithmetical congruence monoid with $\gcd(a, b)$ not a prime power, then $M(a, b)$ is quasi k-furcus for some k.

If S is k-furcus, then whenever an element can be factored into at least k elements, it can be factored as the product of exactly k atoms of S; thus every nonunit has a factorization of length at most k. This yields:

Proposition 3.2. *A k-furcus semigroup S is also quasi k-furcus.*

The converse to Proposition 3.2 is false, as the following example illustrates.

Example 3.3. Consider the arithmetical congruence monoid $S = M(6, 30)$ (in the notation of Example 3.1); its first few elements are $1, 6, 36, 66, 96, 126, 156, \ldots$. From the proof of Theorem 2.3 in [Banister et al. 2007] we know that S is quasi 15-furcus. We will now consider all factorizations of the element 6^{16} in S. In $\mathbb{N}$, 6^{16} factors as $6^{16} = 2^{16}3^{16}$. Since elements of S are multiples of 6 that are congruent to 1 modulo 5, the only factorizations of 6^{16} in S are

$$6^{16}, \quad 96 \cdot 486 \cdot 6^{10}, \quad 1536 \cdot 39366 \cdot 6^{6}, \quad \text{and} \quad 24576 \cdot 3188646 \cdot 6^{2}.$$

Therefore, $\mathcal{L}(6^{16}) = \{4, 8, 12, 16\}$ and S is quasi 15-furcus but not 15-furcus.

In spite of the nonequivalence of k-furcus and quasi k-furcus semigroups, these classes of semigroups share many properties. We illustrate this in Theorem 3.4, which parallels Theorem 2.1 in both statement and proof.

Theorem 3.4. *Let S be a nontrivial quasi k-furcus semigroup and let $0 \neq x \in S$ be a nonunit nonatom. Then*:

(1) *S contains no strong atoms.*

(2) $\rho(S) = \infty$.

(3) $C(S) \leq k+1$.

(4) $3 \leq cr(S) \leq k+1$.

We now give a definition that appears stronger than that of k-furcus, although we have no examples to justify this claim. From Example 2.4, the semigroup with generators $Np_1^{a_1}, Np_2^{a_2}, \dots, Np_k^{a_k}$ has the property that every factorization of an element x with $\mathfrak{L}(x) \geq k$ must have length at least k. This motivates the following definition. We call a semigroup S *strongly k-furcus* if S is k-furcus and no element that can be written as the product of k atoms can be written as the product of less than k atoms. The following theorem gives improvements to Theorem 2.1 when S is strongly k-furcus.

Theorem 3.5. *Let S be a nontrivial strongly k-furcus semigroup*, *and let $0 \neq x \in S$ be a nonunit nonatom. Then*:

(1) $\mathfrak{L}(x) = [k, \sup \mathfrak{L}(x)]$ *or* $\mathfrak{L}(x) \subseteq \{2, 3, \dots, k-1\}$.

(2) $\Delta(x) = \{1\}$ *or* $\Delta(x) \subseteq \{1, 2, \dots, k-3\}$.

(3) $\{1\} \subseteq \Delta(S) \subseteq \{1, 2, \dots, k-3\}$.

We now give an analog to [Adams et al. 2009, Theorem 1.1(1)] when S is strongly k-furcus. This analog was omitted from our Theorem 2.1 since the hypothesis of S being k-furcus but not strongly k-furcus seems not to be enough information to guarantee these results. Again, we point out that we have no example of a k-furcus semigroup that is not also strongly k-furcus.

Proposition 3.6. *If S is a nontrivial strongly k-furcus semigroup and is either left or right cancellative*, *then S contains infinitely many atoms.*

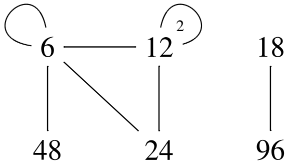
4. Irreducible divisor graphs

In this section we give a means of visually representing the factorization of an element in a k-furcus semigroup. The concept of the irreducible divisor graph of an element in a commutative integral domain was introduced in [Coykendall and Maney 2007] and further studied in [Axtell et al. 2011]. We now give a similar definition for the irreducible divisor graph of an element in a multiplicative semigroup. Given a semigroup S and an element $x \in S$, the *irreducible divisor graph of* x, denoted $G(x)$, is defined as follows. The vertices of $G(x)$ are the atoms $a \in S$ such that $a \mid x$. Two vertices a and b of $G(x)$ are connected by an edge provided $ab \mid x$. Moreover, we place n loops (for $n > 1$ this is denoted by a single loop labeled with an n) on vertex a if $a^n \mid x$ but yet $a^{n+1} \nmid x$. We now provide two examples to illustrate this definition.

Example 4.1. Consider the element $y = 1728$ in the multiplicative bifurcus semigroup $S = 6\mathbb{Z}$. The only factorizations of y in S are

$$\begin{aligned} 1728 &= 6 \cdot 6 \cdot 48 \\ &= 12 \cdot 12 \cdot 12 \\ &= 6 \cdot 12 \cdot 24 \\ &= 18 \cdot 96. \end{aligned}$$

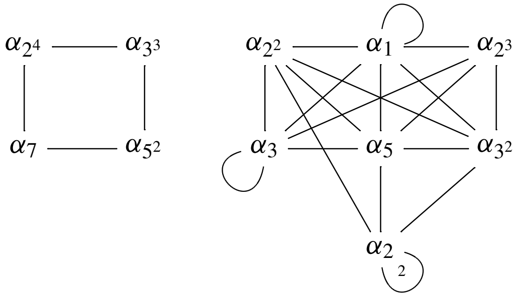
Therefore, $G(y)$, the irreducible divisor graph of y in S is



Example 4.2. Let S be the strongly 4-furcus multiplicative semigroup with generators $210 \cdot 2^{a_1}, 210 \cdot 3^{a_2}, 210 \cdot 5^{a_3}, 210 \cdot 7^{a_4}$, where each a_i is a nonnegative integer given in Example 2.4. Let $x = 2^8 \cdot 3^7 \cdot 5^6 \cdot 7^5 \in S$ and note that x factors only as

$$\begin{aligned} x &= (210 \cdot 2^4)(210 \cdot 3^3)(210 \cdot 5^2)(210 \cdot 7) \\ &= (210 \cdot 2^3)(210 \cdot 3^2)(210 \cdot 5)(210)(210) \\ &= (210 \cdot 2^3)(210 \cdot 3)(210 \cdot 5)(210 \cdot 3)(210) \\ &= (210 \cdot 2^2)(210 \cdot 3^2)(210 \cdot 5)(210 \cdot 2)(210) \\ &= (210 \cdot 2^2)(210 \cdot 3)(210 \cdot 5)(210 \cdot 2)(210 \cdot 3) \\ &= (210 \cdot 2)(210 \cdot 3^2)(210 \cdot 5)(210 \cdot 2)(210 \cdot 2) \end{aligned}$$

Setting $\alpha_{b^j} := 210 \cdot b^j$, the irreducible divisor graph $G(x)$ is



The fundamental result in the theory of irreducible divisor graphs, proved in [Coykendall and Maney 2007; Axtell et al. 2011] tells us that an atomic integral domain R is a UFD if and only if $G(x)$ is connected (equivalently, complete) for all nonunits $x \in R$. In fact, the proof of this result goes through for any commutative, cancellative semigroup. As should be no surprise, the examples above give disconnected graphs. In fact, the following theorem illustrates that this is nearly always the case for strongly k-furcus semigroups, thus giving another demonstration of how "bad" factorization is in k-furcus semigroups.

Theorem 4.3. *Let S be a commutative, cancellative strongly k-furcus semigroup. Then $G(x)$ is disconnected for every nonatom, nonunit x of S with* $\sup \mathcal{L}(x) > k$.

Proof. Divide the set of vertices of $G(x)$ into two subsets:

$$A = \{a \in \mathscr{A}(S) : x = aa_1a_2\cdots a_{k-1};\, a_i \in \mathscr{A}(S)\},$$

containing the vertices involved in factorizations of length k, and

$$B = \{b \in \mathscr{A}(S) : x = bb_1b_2\cdots b_m;\, b_i \in \mathscr{A}(S), m \geq k\},$$

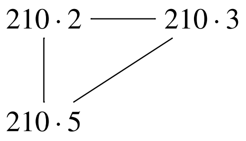
containing those involved in factorizations of length greater than k. Assume $b \in \mathscr{A}(S)$ with $b \in A \cap B$.

Since $b \in A$, $x = bc_1c_2\cdots c_t$, where $t = k-1$. Since $b \in B$, $x = bd_1\cdots d_s$, $s \geq k$. Thus $\frac{x}{b} = c_1c_2\cdots c_t = d_1d_2\cdots d_s$ has a factorization of length greater than or equal to k and a factorization of length less than k, which is impossible since S is strongly k-furcus. Therefore $A \cap B = \varnothing$.

Now assume $a \in A$ and $b \in B$ with an edge connecting a and b in $G(x)$. Then $x = abc_1\cdots c_t$ with c_i atoms of S. If $t = k-2$, then $b \in A$, a contradiction since $b \in B$. If $t > k-2$, then $a \in B$, a contradiction since $a \in A$. Therefore no edges connect vertices in A with vertices in B, and hence $G(x)$ is disconnected. □

The requirement that $\sup \mathcal{L}(x) > k$ is necessary as the following example illustrates.

Example 4.4. Consider the element $x = 2^4 \cdot 3^4 \cdot 5^4 \cdot 7^3$ in the strongly 4-furcus semigroup from Example 4.2 which factors only as $x = (210\cdot 2)(210\cdot 3)(210\cdot 5)$ with $\alpha_{b^j} = 210 \cdot b^j$. Since x has no factorization of length greater than 3, its irreducible divisor graph, shown below is connected.



Acknowledgement

The authors would like to thank Scott Chapman for insightful comments and for the referee's thorough reading of an earlier draft and for several comments that improved the exposition.

References

[Adams et al. 2009] D. Adams, R. Ardila, D. Hannasch, A. Kosh, H. McCarthy, V. Ponomarenko, and R. Rosenbaum, "Bifurcus semigroups and rings", *Involve* **2**:3 (2009), 351–356. MR 2011b:20161 Zbl 1190.20046

[Axtell et al. 2011] M. Axtell, N. R. Baeth, and J. Stickles, "Irreducible divisor graphs and factorization properties of domains", *Comm. Algebra* **39**:11 (2011), 4148–4162. MR 2855118

[Baeth et al. 2011] N. R. Baeth, V. Ponomarenko, D. Adams, R. Ardila, D. Hannasch, A. Kosh, H. McCarthy, and R. Rosenbaum, "Number theory of matrix semigroups", *Linear Algebra Appl.* **434**:3 (2011), 694–711. MR 2012c:11265 Zbl 05833978

[Banister et al. 2007] M. Banister, J. Chaika, S. T. Chapman, and W. Meyerson, "On the arithmetic of arithmetical congruence monoids", *Colloq. Math.* **108**:1 (2007), 105–118. MR 2007m:20096 Zbl 1142.20038

[Coykendall and Maney 2007] J. Coykendall and J. Maney, "Irreducible divisor graphs", *Comm. Algebra* **35**:3 (2007), 885–895. MR 2008a:13001 Zbl 1114.13001

[Geroldinger and Halter-Koch 2006] A. Geroldinger and F. Halter-Koch, *Non-unique factorizations: algebraic, combinatorial and analytic theory*, Pure and Applied Mathematics **278**, CRC, Boca Raton, FL, 2006. MR 2006k:20001 Zbl 1113.11002

Received: 2011-06-23 Revised: 2012-02-07 Accepted: 2012-02-09

baeth@ucmo.edu
Department of Mathematics and Computer Science, University of Central Missouri, Warrensburg, MO 64093, United States

cassity@ucmo.edu
Department of Mathematics and Computer Science, University of Central Missouri, Warrensburg, MO 64093, United States

INVOLVE 5:3 (2012)
dx.doi.org/10.2140/involve.2012.5.303

Studying the impacts of changing climate on the Finger Lakes wine industry

Brian McGauvran and Thomas J. Pfaff

(Communicated by Robert B. Lund)

We report the results of a project with Six Mile Creek Winery in Ithaca, NY, in which we investigate possible climate impacts on the area wine industry. Specifically, we examine winter minimum temperatures in Ithaca since temperatures below −15°F damage buds of French-American hybrid grapes and temperatures below −5°F affect Vinifera grape varieties. We used the generalized extreme value distribution to model the winter minimum and adjusted this model based on climate simulation data.

1. Introduction

Climate change is a serious issue facing society and it is now feasible to address local questions concerning climate. Simulations are now being run using a 50 km^2 grid; meaning there are data streams for each 50×50 km box covering the United States that can be analyzed to address specific local questions about future climate. These data streams provide data every three hours consisting of surface specific humidity, precipitation, surface pressure, surface downwelling shortwave radiation, surface air temperature, zonal surface wind speed, and meridional surface wind speed for the periods from 1968 to 2000 and 2038 to 2070. The reason for the two periods is that computer simulations do not necessarily predict future climate well (note that climate is considered to be a distribution resulting from approximately 30 years of weather) but they are consistent in that the changes in variables from the recent scenario 1968–2000 to the future scenario 2038–2070 represent an estimate of the change in these variables over those two periods. For example, the sample average winter minimum temperature in the computer simulation for 1968–2000 is −19.25°F, which is a few degrees off from the sample average winter minimum temperature for that period of −16.09°F. The sample average winter minimum from the computer simulation for the future scenario of 2038–2070 is −12.92°F.

MSC2010: 62G05, 62G32.

Keywords: generalized extreme value distribution, winter minimum temperatures, climate change.

What this suggests to us is that the mean winter minimum temperature is likely to increase 6.33°F to around −9.76°F.

We chose to study changes in the minimum winter temperature in our area (the Finger Lakes region is a major wine producing region) because the French-American hybrid grapes will begin to have bud damage when the temperature falls below −15°F, whereas Vinifera grape varieties (such as Riesling, Cabernet Franc, and Chardonnay) will start to have bud damage when temperatures fall below −5°F. We note that all of these grapes are currently being grown despite cold winter temperatures, with the Riesling grape being a premier product, due to microclimates near the Finger Lakes and various techniques to keep the vines warmer in the winter. Our goal is to estimate the probability that the winter minimum temperature in Ithaca, NY, will fall below −15°F and will fall below −5°F by mid-century. To do this we need to estimate the current distribution of winter minimum temperatures and use the simulations to estimate how this distribution may change in the future.

2. Winter minimum distributions

Since winter minimum temperatures are extreme events, an appropriate distribution to use to model the data is the generalized extreme value (GEV) family for minima given by

$$G(z) = 1 - \exp\left(-\left(1 - \xi\,\frac{z-\mu_t}{\sigma}\right)^{-1/\xi}\right) \tag{1}$$

defined on $\{z : 1 - \xi(z-\mu_t)/\sigma > 0\}$, where $-\infty < \mu_t, \xi < \infty$ and $\sigma > 0$. Here μ_t is the location parameter, σ is the scale parameter, and ξ is the shape parameter. For this paper we used the Extremes Toolkit [Katz et al. 2009] for the statistical software R to fit the GEV model to our data sets. The Extremes Toolkit is maximizing the log-likelihood function

$$l(\mu_t, \sigma, \xi) \\ = -m\log\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\log\left(1+\xi\,\frac{z_i-\mu_t}{\sigma}\right) - \sum_{i=1}^{m}\left(1+\xi\,\frac{z_i-\mu_t}{\sigma}\right)^{-1/\xi},$$

under the conditions that $\xi \neq 0$ and $1+\xi\,\dfrac{z_i-\mu_t}{\sigma} > 0$ for all i. When $\xi = 0$, the log-likelihood function to be maximized is

$$l(\mu_t, \sigma) = -m\log\sigma - \sum_{i=1}^{m}\frac{z_i-\mu_t}{\sigma} - \sum_{i=1}^{m}\exp\left(-\frac{z_i-\mu_t}{\sigma}\right).$$

In either case this yields the maximum likelihood estimate for the parameter vector (μ_t, σ, ξ). The Extremes Toolkit also allows for any of the three variables to be time varying along with indicator variables, and we will use a location parameter of the form $\mu_t = \mu_0 + \alpha t + \beta 1_{t>t_0}$ ($1_{t>t_0} = 1$ if $t > t_0$ and 0 if $t \le t_0$). Coles

1893	−14	1914	−9	1943	−10	1964	−23	1986	−18
1894	−9	1915	−8	1944	−14	1965	−14	1987	−14
1895	−18	1916	−11	1945	−8	1966	−14	1988	−19
1896	−6	1917	−22	1946	−9	1967	−23	1989	−15
1897	−4	1926	−11	1947	−20	1968	−7	1990	−7
1898	−16	1927	−2	1948	−5	1969	−17	1991	−7
1899	−2	1928	−7	1949	−8	1970	−11	1992	−17
1900	−8	1929	−8	1950	−7	1971	−11	1993	−24
1901	−2	1930	−1	1951	−10	1972	−17	1994	−12
1902	−5	1931	3	1952	2	1973	−12	1995	−16
1903	−20	1932	−1	1953	−14	1974	−12	1996	−11
1904	−6	1933	−24	1954	−13	1975	−21	1998	−13
1905	−11	1934	−11	1955	−7	1976	−17	1999	−11
1906	−10	1935	−6	1956	−25	1978	−23	2000	−5
1907	−13	1936	−5	1957	−14	1979	−13	2001	6
1908	−8	1937	−9	1958	−10	1980	−21	2002	−14
1909	−6	1938	−2	1959	−10	1981	−23	2003	−17
1910	−1	1939	−3	1960	−25	1982	−15	2004	−22
1911	−16	1940	−3	1961	−17	1983	−22	2005	−11
1912	−1	1941	−8	1962	−18	1984	−11	2006	−9
1913	−15	1942	−14	1963	−11	1985	−8	2007	−5

Table 1. Observed minimum winter temperature for the years 1893 to 2007, in °F. Data in all tables and figures refer to Ithaca, NY.

[2001] discusses the GEV distribution and modeling details. We obtained the winter minimum temperatures for Ithaca from the NOAA (National Oceanic and Atmospheric Administration) National Data Center [NNDC 2009] for the years 1893 through 2008, but winter minimums for 1977 and 1997 were missing. The observed data is listed in Table 1.

3. Results

The parameters for the GEV fitted to the observed data are in Table 2, for the model $\mu_t = \mu_0 + \alpha t$ of the location parameter, where t is scaled so that $t = 0$ for 1893 and $t = 1$ for 2007. The time scale was significant ($p = 0.004$) with a coefficient of -6.26649°F. This seems rather large and we conjecture that this is due to the colder temperatures in the 1960s and the fact that we have more data before 1960 than after. A scatter plot of the data, Figure 1, does not show any clear changepoints, although the winter minimum temperatures appear lower from roughly 1960–1980.

μ_0	-6.07284	(1.28365)
α	-6.26649	(2.12176)
σ	6.47464	(0.48768)
ξ	-0.31045	(0.06412)

Table 2. Expected values and standard errors of parameters for observed Ithaca winter minimum temperatures for the winters 1893–2007 with the location parameter $\mu_t = \mu_0 + \alpha t$; here t is scaled so that $t = 0$ for 1893 and $t = 1$ for 2007.

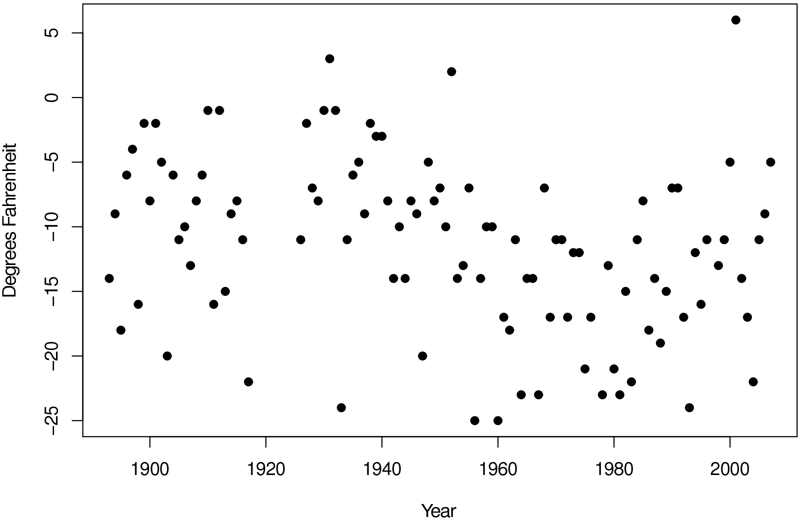


Figure 1. Observed winter minimum temperatures.

There are minor weather station changes in 1969 and 1987, and it may be that since the station is associated with Cornell University that the station is particularly stable. In 1969 the station was raised 10 feet and the longitude changed from −76.466670° to −76.45000°. In 1987 there was a change in equipment. Despite these seemingly minor changes there appear to be some changepoint issues. We first used a model with

$$\mu_t = \begin{cases} \mu_0 + \alpha t & \text{for } t < 1969, \\ \mu_0 + \alpha t + \beta_1 & \text{for } 1969 \le t \le 1987, \\ \mu_0 + \alpha t + \beta_2 & \text{for } t > 1987, \end{cases} \tag{2}$$

to test all changepoints simultaneously. A likelihood ratio test with the Extremes Toolkit was used to check for significant differences between the model with and without changepoints. The test uses the deviance statistic $D = 2\{\ell_1(\mathcal{M}_1) - \ell_0(\mathcal{M}_0)\}$,

μ_0	-4.91412	(1.39459)
α	-10.10339	(2.76638)
β	4.48365	(2.09895)
σ	6.27504	(0.47723)
ξ	-0.29469	(0.06717)

Table 3. Expected values and standard errors (in parentheses) of parameters for observed winter minimum temperatures for the winters 1893–2007 with the location parameter μ_t given by (4) and t scaled so that $t = 0$ for 1893 and $t = 1$ for 2007.

where $\mathcal{M}_0$ is a submodel of $\mathcal{M}_1$ and $\ell_0(\mathcal{M}_0)$ and $\ell_1(\mathcal{M}_1)$ are the maximized values of the log-likelihood for their respective models [Coles 2001, p. 35]. The test was not significant ($p = 0.0514$) and so the addition of the changepoints does not improve the model. We also tested the changepoints individually by using the following expressions for μ_t:

$$\mu_t = \begin{cases} \mu_0 + \alpha t & \text{for } t < 1969, \\ \mu_0 + \alpha t + \beta & \text{for } t \geq 1969, \end{cases} \tag{3}$$

$$\mu_t = \begin{cases} \mu_0 + \alpha t & \text{for } t \leq 1987, \\ \mu_0 + \alpha t + \beta & \text{for } t > 1987. \end{cases} \tag{4}$$

In testing each of these against the model without any changepoints the model with (3) was not significant ($p = 0.6951$) and the model with (4) was significant ($p = 0.0341$). We will use the model with μ_t given by (4). The values of the parameters for the observed model we will use with μ_t given by (4) are given in Table 3.

The climate simulations used are the GFDL RCM3 data found at [NARCCAP 2009]. The future scenario simulations assume the IPCC A2 scenario for greenhouse gas emissions. In brief, the A2 scenario is an estimate of our future greenhouse gas emissions based on socioeconomic factors. It assumes our emissions will continue to rise, but at a decreasing rate. The scenario predicts these emissions based on the idea that the global economy will become more regional and each region will become more self-reliant. These various regions are less involved internationally. Similarly global environmental concerns are weak, and regional attempts at controlling pollution are only enough to maintain their environmental amenities [IPCC 2001].

The simulation data we used is listed in Table 4. The simulation data is for the winter periods of 1968–1999 and 2038–2069. Here we take the approach that each period has its own GEV distribution, since climate is considered to be 30 years of weather, and we will test to see if there are any statistically significant changes in

1968	−21.6	1984	−13.6	2038	−19.7	2054	−5.0
1969	−19.1	1985	−25.8	2039	−14.3	2055	−11.5
1970	−12.8	1986	−22.8	2040	−19.9	2056	−15.4
1971	−24.8	1987	−15.0	2041	−6.1	2057	−7.2
1972	−24.0	1988	−11.8	2042	−14.4	2058	0.3
1973	−18.6	1989	−15.3	2043	−18.9	2059	−10.8
1974	−12.0	1990	−6.7	2044	−17.6	2060	−4.8
1975	−12.7	1991	−13.7	2045	−8.8	2061	−16.9
1976	−23.0	1992	−20.6	2046	−19.1	2062	−3.8
1977	−14.9	1993	−11.9	2047	−8.4	2063	−15.4
1978	−10.9	1994	−14.9	2048	−5.6	2064	−5.1
1979	−20.3	1995	−16.6	2049	−4.8	2065	−3.3
1980	−18.2	1996	−20.9	2050	−10.0	2066	−9.9
1981	−27.6	1997	−19.2	2051	−14.7	2067	−22.4
1982	−22.3	1998	−21.0	2052	−3.4	2068	−10.4
1983	−15.0	1999	−14.3	2053	−20.0	2069	−9.4

Table 4. Simulated minimum winter temperatures, using the IPCC A2 scenario for greenhouse gas emissions, obtained from the GFDL RCM3 simulations found at [NARCCAP 2009].

those parameters. We assume that the simulated data will provide a good estimate of any changes in the parameters under the A2 scenario, even if the simulations do not match observation well, as noted in the introduction. If there is a significant change in any parameter, then we will apply that change to the observed parameter, yielding a future distribution. To detect significant differences in the parameters between the simulated winter minimums for the winters of 1968–1999 and 2038–2069 we fit a GEV distribution to all the simulated data with indicator functions for each parameter. For example, the location parameter is of the form $\mu_t = \mu_0 + \gamma 1_{t>2000}$ ($1_{t>2000} = 1$ if $t > 2000$ and 0 if $t \leq 2000$) and if the addition of the indicator variable is significant using the likelihood ratio test, then the coefficient of the indicator provides an estimate of how much to adjust the observed parameter. The only significant indicator variable is for the location parameter and has a coefficient of 6.33435°F ($p < 0.001$). In other words, we expect a shift of 6.33435°F in the location parameter from the period 1968–1999 to 2038–2069. We do not expect a change in the scale or shape parameters. We will shift over the same time period and hence we will take our observed location parameter to be

$$\mu = -4.91412 - 10.10339(0.929825) + 4.48365(1) = -9.82485°\text{F}.$$

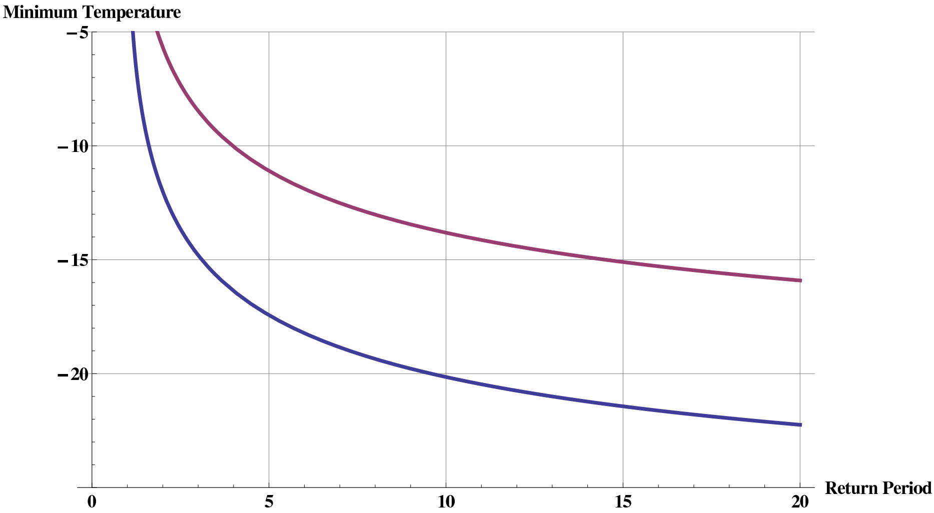


Figure 2. Return level curves for the observed winter minimum ($t = 0.9298$, winter of 1999), bottom curve, and the predicted mid-century winter minimum.

(Note the 1999 winter minimum corresponds to $t = 0.929825$ since t was scaled to start in 1893, $t = 0$, and end in 2007, $t = 1$). Hence our predicted location parameter for mid-century is

$$\mu = -9.82485 + 6.33435 = -3.4905°\text{F}.$$

We will use the same scale and shape parameters from our observed data. In summary, the value of the simulation data is that it estimates a shift in the location parameter for winter minimum temperatures under an IPCC A2 scenario of 6.33435°F from the period 1968–1999 to 2038–2069. Now, while our observed location parameter does have a time variable extrapolating this parameter to mid-century would be too far of an extrapolation, and would not necessarily take into consideration changing greenhouse gasses as the simulation data does. In fact, we noted earlier that the coefficient of the time parameter may just reflect a decade of cooler temperatures. We should also mention that there are statistical downscaling methods to build models that combine the observed and simulated data to make projections [Wilby et al. 2004].

Two return level curves, observed and predicted, are plotted in Figure 2. In these graphs we see, for example, that the observed data tells us that about once every 10 years the winter minimum temperature will fall below −20°F. On the other hand, we predict that by mid-century the winter minimum temperature will fall below −14°F only once every 10 years.

Temperature	Probability below temperature	
	Observed data	Predicted climate
−25°F	1.44%	0.00%
−20°F	10.45%	0.63%
−15°F	32.21%	6.90%
−10°F	62.18%	25.17%
−5°F	86.46%	54.12%
0°F	97.33%	83.25%

Table 5. The probabilities that the winter minimum temperature will fall below the given values for the observed winter minimum and predicted winter minimum.

Table 5 gives more detailed information about the probability that the winter minimum temperature will fall below a given temperature. Focusing on the −15°F threshold, we see that currently there is a 32% chance the temperature will drop below −15°F but by mid-century this drops to only a 7% chance. Similarly, the chance of dropping below −5°F changes from 86% to 54%.

4. Conclusion and discussion

The changes in the ability of the Finger Lakes wine industry to grow French-American hybrid grapes and Vinifera grapes is still unclear. For one, we still need to consider area microclimates. The weather station from which we collected data is within the region, but it is not located on the shore of any of the Finger Lakes. Land that is adjacent to a lake, in particular Cayuga and Seneca, have warmer winter temperatures. For instance, suppose your winery location was typically 5°F warmer in the winter than the weather station location due to being in a lake microclimate. By looking at the −20°F row in Table 5, the chances of a minimum below −15°F falls from about 1 in 10 years to about 1 in 100 years. We also point out that we are using a general −15°F and −5°F cutoffs when, in fact, there are varietal differences.

We see this study as the beginning of aiding the Finger Lakes wine industry in dealing with climate instability. While this study may seem to suggest a positive effect for the region, there are other possible impacts that are likely to be negative that still need to be studied. For example, in the winter of 2004 the industry experienced a single freeze event in January that reduced the *Vitis vinifera* crop by almost half [Zabadal et al. 2007]. In general, grapevines are susceptible to rapid temperature drops during the acclimation and deacclimation periods of dormancy. Also, increasing rain or humidity would increase mildew problems, which would increase the need for mildew control. Finally, the region is known for its Riesling wines but the riesling grapes require an adequate number of cool nights at the end of the growing season. If there were some change in the cool night numbers that reduced the quality of the Riesling wine, then the region would have to rebrand itself with some other wine. On the other hand, it is also worth studying how much the summer growing season may change as longer warmer summers would help the industry.

Acknowledgments

We wish to thank the North American Regional Climate Change Assessment Program (NARCCAP) for providing the data used in this paper. We also thank Doug Nychka and Seth McGinnis from the National Center for Atmospheric Research (NCAR), Eric Russo and Paul King from Six Mile Creek Vineyard, and the referees of this paper.

References

[Coles 2001] S. Coles, *An introduction to statistical modeling of extreme values*, Springer, London, 2001. MR 2003h:62002 Zbl 0980.62043

[IPCC 2001] "IPCC special report on emissions scenarios", Report, Intergovernmental Panel on Climate Change, 2001, available at http://www.grida.no/publications/other/ipcc_sr/?src=/climate/ipcc/emission/094.htm.

[Katz et al. 2009] R. Katz, B. Brown, E. Gilleland, and D. Nychka, "The weather and climate impact assessment science program: extremes toolkit", Statistical software, National Center for Atmospheric Research, 2009, available at http://www.assessment.ucar.edu/toolkit.

[NARCCAP 2009] "Data", data set, North American Regional Climate Change Assessment Program, 2009, available at http://narccap.ucar.edu/data/index.html.

[NNDC 2009] "Surface climate data", data set, NOAA National Data Centers, 2009, available at http://cdo.ncdc.noaa.gov/qclcd/QCLCD?prior=N.

[Wilby et al. 2004] R. L. Wilby, S. P. Charles, E. Zorita, B. Timbal, P. Whetton, and L. O. Mearns, "Guidelines for use of climate scenarios developed from statistical downscaling methods", Report, IPCC Task Group on Data and Scenario Support for Impact and Climate Analysis (TGICA), 2004, available at http://www.ipcc-data.org/guidelines/dgm_no2_v1_09_2004.pdf.

[Zabadal et al. 2007] T. J. Zabadal, M. C. Goffinet, M. L. Chien, I. Dami, and T. E. Martinson, *Winter injury to grapevines and methods of protection*, MSU Extension Bulletin **E2930**, Michigan State University Extension, East Lansing, MI, 2007.

Received: 2011-06-23 Revised: 2011-11-21 Accepted: 2011-11-23

brian.mcgauvran@gmail.com *27 Sandy Hill Rd, Commack, NY 11725, United States*

tpfaff@ithaca.edu *Mathematics Department, Ithaca College, Ithaca, NY 14850, United States*

INVOLVE 5:3 (2012)
dx.doi.org/10.2140/involve.2012.5.313

A graph-theoretical approach to solving Scramble Squares puzzles

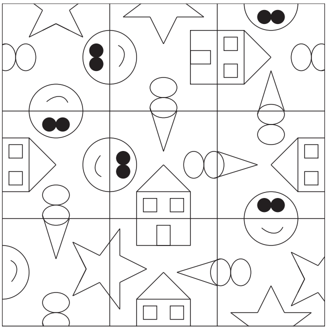
Sarah Mason and Mali Zhang

(Communicated by Arthur T. Benjamin)

A Scramble Squares puzzle is made up of nine square pieces such that each edge of each piece contains half of an image. A solution to the puzzle is obtained when the pieces are arranged in a 3×3 grid so that the adjacent edges of different pieces together make up a complete image. We describe a graph-theoretical approach to solving Scramble Squares puzzles and a method for decreasing randomness in the backtracking solution algorithm.

1. Introduction

A Scramble Squares® puzzle (created and marketed by B. Dazzle, Inc.) consists of nine square pieces, each of which contains half of an image on each side. A solution to a Scramble Squares puzzle is an arrangement of the nine pieces into a 3×3 grid so that the adjacent half images on adjacent pieces together create a complete image. Here is an example of a solution to a Scramble Squares puzzle:



There are many different ways to arrange the pieces in an attempt to solve a Scramble Squares puzzle. There are nine different positions in the 3×3 grid and therefore 9! different ways to place the pieces into the grid, assuming that the pieces are pairwise distinct. Once the pieces have been placed, there are 4

MSC2010: 05C75, 94C15.
Keywords: graph theory, algorithms.

different orientations for each piece. This means that there are a total of $4^9 \times 9!$ different arrangements of the pieces. Taking into account rotational symmetry, if there is a solution there must be at least four, but still the probability of finding one of them by laying the pieces down at random can be as low as $4/(4^9 \times 9!)$, or about 4.2×10^{-11}. It would therefore be desirable to have an efficient algorithm for solving Scramble Squares puzzles, but this turns out to be quite a steep request since Scramble Squares are *constraint satisfaction problems* (CSPs) and many CSPs are known to belong to the NP-complete complexity class. The most efficient known algorithm for solving Scramble Squares puzzles is a depth first backtracking search developed by Brandt, Burger, Downing, and Kilzer [Brandt et al. 2002].

A visual representation of a problem can often provide key insights into the nature of the solution(s). The graph-theoretical solution to the Instant Insanity puzzle is a wonderful example of this phenomenon [Busacker and Saaty 1965; Carteblanche 1947; Grecos and Gibberd 1971; Van Deventer 1969]. The Instant Insanity puzzle consists of four unit cubes whose faces are colored arbitrarily with four colors. A solution is obtained by stacking the cubes into a vertical rectangular prism with dimensions $4 \times 1 \times 1$ so that each color appears exactly once on each side of the prism. Van Carteblanche [1947] introduces a method (elaborated upon by many [Busacker and Saaty 1965; Grecos and Gibberd 1971; Van Deventer 1969]) for representing the cubes as edges in a graph whose vertices correspond to the four colors. A solution is determined by choosing an appropriate subgraph. This graph-theoretical solution to Instant Insanity is the inspiration for this paper. We provide a graph-theoretical solution to a simplified Scramble Squares puzzle, following a similar approach. We also provide a method for ordering the pieces used in the backtracking algorithm in [Brandt et al. 2002] as a way to potentially improve upon its efficiency.

2. Restricted Scramble Squares puzzles

We begin by introducing the terminology and notations which will appear throughout this paper. A *pattern* is a complete image in the puzzle. Each pattern is comprised of two *pictures*, which are halves of the image. The *complement* of a picture is the other half of the pattern. A *piece* is one of the nine squares that make up a puzzle. See Figure 1 for an example.

In this section, we will restrict to puzzles containing four or fewer patterns. We do this for the sake of simplicity, but it would not be difficult to extend these results to puzzles with more patterns; in essence, it involves considering graphs with more vertices but whose solution graphs satisfy the same set of restrictions.

The recording graph. We provide a method to represent any Scramble Squares puzzle mathematically as a graph. Begin by assigning a number to each pattern.

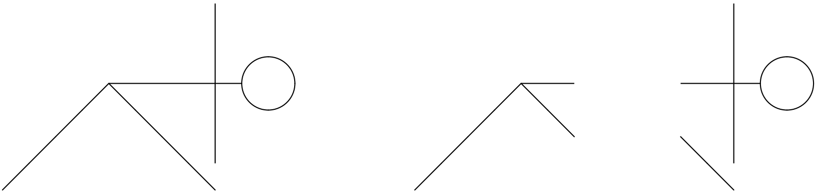


Figure 1. The two pictures on the right are *complements*, which together make up the *pattern* on the left.

Each pattern consists of two pictures, so associate a plus sign to one of the pictures and a minus sign to the other. This assigns a number and a sign to each picture appearing in the puzzle. Notice that two pictures with the same number but opposite signs together form a complete pattern. Also note that if $X+$ is the signed number corresponding to a given picture (of pattern $|X|$), then its complement X^c, is given by $X-$. For example, in Figure 2 the number 1 represents the star, the number 2 represents the ice cream cone, the number 3 represents the house, and the number 4 represents the smiling face. For this reason, we use the absolute value notation to denote the underlying pattern, so that $|X+| = |X-|$, and we frequently refer to a pair of complementary pictures as X and X^c.

A *repetition* in a puzzle piece is a picture which appears more than one time on the piece. Note that a picture $X+$ and its complement $X-$ appearing on the same piece do not constitute a repetition. We say that a puzzle is *repetition-free* if no piece of the puzzle contains a repetition. This means that a particular picture may appear multiple times in the puzzle, provided that each appearance is on a different piece. We restrict to 2 × 2 repetition-free puzzles but it would be interesting to extend these results to larger puzzles or puzzles containing repetitions. See Section 4 for details on this and other related open problems.

We construct a graph, called the *recording graph* $G(P)$, corresponding to a given Scramble Squares puzzle P as follows. The vertices of $G(P)$ are the symbols associated to the pictures appearing in the puzzle pieces. They are arranged into

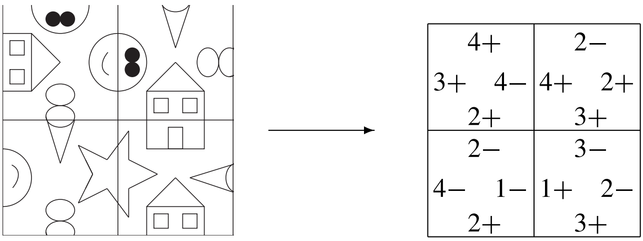


Figure 2. Converting pictures to symbols.

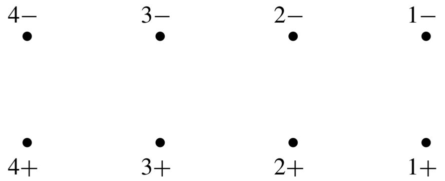


Figure 3. An edgeless graph.

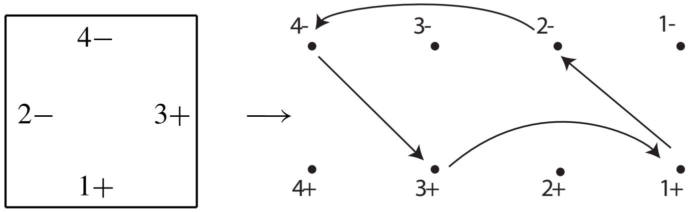


Figure 4. One piece of the puzzle becomes a length-4 directed cycle.

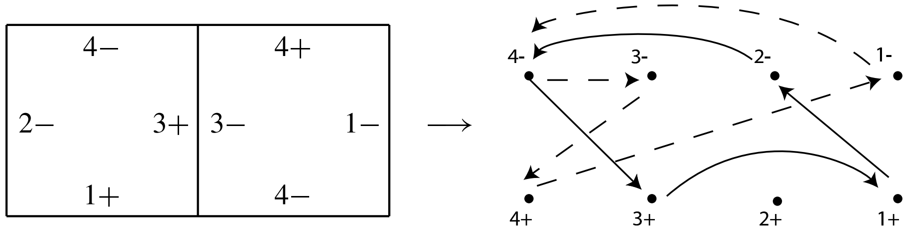


Figure 5. The left-hand piece is represented by solid lines, while the right-hand piece is represented by dashed lines.

two rows so that the top row contains the pictures with negative sign and the bottom row contains the pictures with positive sign. The vertices are written in decreasing order in both rows, as shown in Figure 3.

The edges of the recording graph are colored directed edges obtained from the pieces in the puzzle. Each piece is assigned a color. (Note that the numbers represent patterns while the colors represent pieces.) Construct four directed edges for each piece by drawing an arrow from each picture appearing in the piece to the picture which is ninety degrees away clockwise. Therefore each piece contributes four edges to the recording graph. The vertex from which this arrow originates is called the *tail* of the edge, while the vertex to which it points is called the *head* of the edge. Figure 4 demonstrates the construction of the four edges corresponding to one puzzle piece, and Figure 5 demonstrates the recording graph for a Scramble Squares puzzle with two pieces.

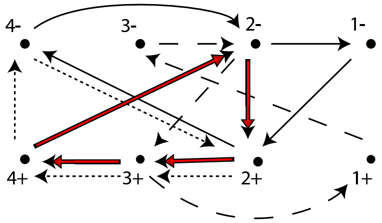


Figure 6. Recording graph for the puzzle shown in Figure 2.

The pieces of the puzzle are distinguished from one another by the color (or shading) of their edges. Once all the pieces have been represented in the graph, the resulting figure is called the *recording graph*. Figure 6 shows an example. We may now discard the original pieces since the recording graph encodes all of the information necessary to solve the puzzle. We determine a solution by finding a subgraph of the recording graph which satisfies certain properties.

Solution graphs for* 2×2 *repetition-free puzzles. Every solution to a 2×2 Scramble Squares puzzle without repetitions is an arrangement of the pieces such that each picture not on the boundary is adjacent to its complement. Every subgraph of the recording graph which contains four edges of distinct colors represents an arrangement of the pieces. (Note that we need exactly one edge of each color to represent an arrangement of the pieces since each color represents a piece.) Recall that an arrow $A \to B$ in the recording graph represents the corner between sides A and B, where A is 90 degrees counterclockwise from B. When that edge is present in a subgraph, it means that this corner will be the corner of that piece which is in the middle of the arrangement, adjacent to the other pieces. Since not every arrangement of the pieces constitutes a solution, not every four-colored subgraph of the recording graph constitutes a solution. See Figure 6 for an example of a recording graph, and Figure 7 for examples of a solution subgraph and a subgraph which does not correspond to a solution. We provide necessary and sufficient conditions on a subgraph to guarantee that it constitutes a solution.

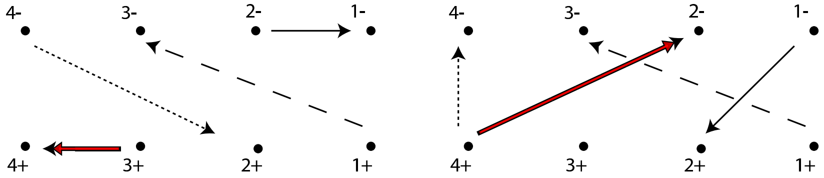


Figure 7. Left: a subgraph of Figure 6 representing the solution shown in Figure 2. Right: graph of an arrangement of the pieces that does not constitute a solution.

In order to state these conditions, we need the notion of *pseudoconnectedness*. Two distinct connected components of a recording graph are said to be *pseudoconnected* if the intersection of the set of absolute values of their vertices is nonempty. Write $C_1 \simeq C_2$ if C_1 and C_2 are pseudoconnected. A *pseudo-path* between two connected components C and D is a collection of connected components $\{C_0 = C, C_1, \dots, C_k = D\}$ such that $C_0 \simeq C_1 \simeq \dots \simeq C_k$. A subgraph of a recording graph is said to be *pseudoconnected* if there is a pseudo-path between every pair of connected components in the graph.

For example, let C_1, C_2, C_3 be the three connected components of a recording graph G, with respective vertex sets $\{1+, 3-, 4-\}$, $\{2+, 3-, 3+\}$, and $\{1-, 4+\}$. The graph G is pseudoconnected even though C_2 is not pseudoconnected to C_3, since there exists a pseudo-path $C_2 \simeq C_1 \simeq C_3$.

Theorem 2.1. *A subgraph of the recording graph $G(P)$ consisting of four edges is a solution graph $G_s(P)$ for a repetition-free 2×2 puzzle if and only if it is a pseudoconnected subgraph satisfying the following properties*:

(1) *Each edge is a different color.*

(2) *The in-degree of each vertex is equal to the out-degree of its complement.*

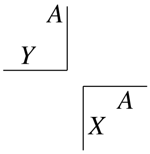
(3) *If $X \to A \to Y$ is a directed path in $G_s(P)$, then Y must be the complement of X.*

Proof. We begin by proving that every subgraph which corresponds to a solution must be of the form described in Theorem 2.1. The subgraph must contain exactly four distinctly colored edges since a solution must use each of the four pieces.

Next consider the pseudoconnectedness property. In a solution to the puzzle, every pair of pieces is either adjacent or diagonally opposite one another. If two pieces are adjacent, then their corresponding vertices are pseudoconnected in the solution graph since the adjacent edges of the pieces must contain the same pattern. If two pieces are diagonally opposite one another, there is a piece between them whose edges share a pattern with each. The edges of this piece will form a pseudo-path between the two corresponding patterns contained in the diagonally opposite pieces. Hence every solution graph must be pseudoconnected.

Every vertex appearing in the solution graph corresponds to a picture which is matched to its complement. Since at every such matching, one picture is represented by the head of a directed edge and the other is represented by the tail of a directed edge, the matching contributes 1 to the in-degree of one picture and 1 to the out-degree of its complement. Therefore the in-degree of a vertex must equal the out-degree of its complement.

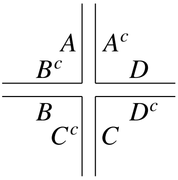
Finally consider a graph that does not satisfy the third property. This implies that the graph contains a length 2 directed path $X \to A \to Y$ such that Y is not the complement of X. Without loss of generality, let the corner $A \to Y$ be the upper-left corner of the solution. Then the corner represented by $X \to A$ cannot be placed in a position adjacent to the corner represented by $A \to Y$ since the complement of A is not A and the complement of Y is not X. Therefore the corner represented by $X \to A$ must be positioned in the lower right-hand corner (diagonally opposite the corner represented by $A \to Y$):



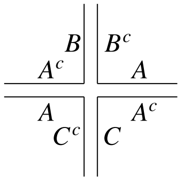
In this case, the picture A^c must appear twice in the piece located in the upper right-hand corner. This contradicts the assumption that the puzzle is repetition-free, and therefore in this case no solution exists. Thus if there is a length two directed path $X \to A \to Y$, then Y must be the complement of X. This implies that the conditions listed in Theorem 2.1 are necessary.

Next we must prove that all subgraphs satisfying the given conditions are indeed solution graphs. Let G be a subgraph satisfying the hypotheses of Theorem 2.1. We prove that the pieces represented by G constitute a solution to the puzzle.

First assume that four distinct patterns appear in the pieces represented by G. Without loss of generality let A, B, C, and D denote the pictures with out-degree one. We can't have $X \to X^c$ for any X by the pseudoconnectedness condition. (Since four distinct patterns appear, if we had $X \to X^c$ then the pattern represented by X would not be pseudoconnected to any of the other patterns, violating pseudoconnectedness.) Therefore without loss of generality assume $A \to B^c$ is one of the pieces. If $B \to A^c$ is a piece, then pseudoconnectedness fails since the patterns $|A|$ and $|B|$ would not be pseudoconnected to the patterns $|C|$ and $|D|$. So, again without loss of generality, assume $B \to C^c$. Then $C \to D^c$ since pattern D must be pseudoconnected to one of the other patterns and if $C \to A^c$ then pattern D would be isolated. Then $D \to A^c$ by process of elimination. Therefore a solution is given by the following arrangement.

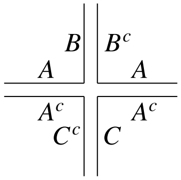


Next assume that three distinct patterns appear in the pieces represented by the subgraph G. Let $|A|$ be the repeated pattern. Then the tails are either given by A, A, B, C or A, A^c, B, C for some pictures A, B, C. Assume that the tails are A, A, B, C. By pseudoconnectedness one of B, C must appear on the same piece as A^c. Assume without loss of generality that this piece is $B \to A^c$. Then C appears (also by pseudoconnectedness) on the same piece as either A^c or B^c. If $C \to A^c$, then the other pieces must be $A \to B^c$ and $A \to C^c$ and a solution is given below.



If $C \to B^c$, then the other two directed edges appearing in G must be $A \to C^c$ and $A \to A^c$, which together with $C \to B^c$ and $B \to A^c$ represent a solution.

If the tails are A, A^c, B, C, then one of B or C must be on the same piece as either A or A^c. Without loss of generality assume this piece is $B \to A$. Then the third condition implies that $A \to B^c$ is another piece. Pseudoconnectedness implies that the remaining two pieces are represented by $C \to A^c$ and $A^c \to C^c$ since $C \to C^c$ would isolate pattern C, keeping it from being pseudoconnected to A or B. Therefore a solution is obtained by placing the pieces as shown below.



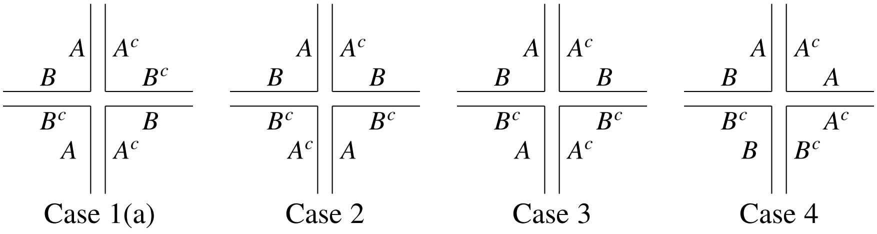
Next assume that two distinct patterns appear in the pieces represented by G. This can happen with each pattern appearing twice or one pattern repeated three times. If each of two patterns $|A|$ and $|B|$ is repeated twice, we may assume without loss of generality that $A \to B$ appears in G. Condition 3 implies that none of $A^c \to A$, $B \to B^c$, or $B \to A$ appears in G, but there must be at least one more piece involving both $|A|$ and $|B|$ since each pattern occurs an even number of times. This piece could be any of $B^c \to A$, $B^c \to A^c$, $B \to A^c$, $A \to B$, $A \to B^c$, $A^c \to B$, or $A^c \to B^c$.

Case 1: If this piece is $B^c \to A$ then A^c must appear at least two more times, once as a head and once as a tail by condition 2. Similarly, B must appear as a head and B^c as a tail by condition 2. This means that the other pieces appearing are either

$$\text{(a)}\ \ A^c \to B \text{ and } B^c \to A^c, \qquad \text{or} \qquad \text{(b)}\ \ A^c \to B^c \text{ and } B \to A^c.$$

In both cases, a solution is possible. See Figure 8(a) for the solution to Case 1(a); Case 1(b) is similar. (Notice that if $B^c \to A$ is replaced by $B \to A^c$ then the proof that the puzzle has a solution is the same.)

Case 2: If the second piece involving $|A|$ and $|B|$ is $A^c \to B$, then the other two pieces must be $B^c \to A$ and $B^c \to A^c$, which are the same pieces used in Case 1(a); see Figure 8, left. A similar argument works when the second piece is $A \to B^c$; see second diagram in Figure 8.



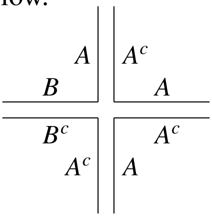
Case 1(a) Case 2 Case 3 Case 4

Figure 8. Solutions for puzzles with two repeated patterns.

Case 3: If the second piece is $A^c \to B^c$ then the other pieces must be $B^c \to A$ and $B \to A^c$ by conditions two and three, which together represent a solution depicted in the third diagram of Figure 8. If the second piece is $A \to B$ then the other two pieces are $B^c \to A^c$, which together with the first two pieces represent a solution similar to the solution for the puzzle with second piece $A^c \to B^c$.

Case 4: Finally, if the second piece is $B^c \to A^c$ then the other two pieces are either ($A \to A^c$ and $B^c \to B$) or ($A \to B$ and $B^c \to A^c$) or ($B^c \to A$ and $A^c \to B$) or ($A \to B^c$ and $B \to A^c$), all of which admit a solution similar to the previous solutions; the solution to the first is depicted in the rightmost part of Figure 8.

Finally, suppose that one of the patterns appears three times. Then we may assume $A \to B$ is a piece, since the two patterns A and B must be pseudoconnected. Assuming A is the piece repeated three times, there is one piece containing B^c as the tail and either A or A^c as the head. Since the remaining two pieces must contain two occurrences of A and two occurrences of A^c by the repetition-free assumption, this second piece must be $B^c \to A^c$. The remaining pieces must both be $A \to A^c$ since $A^c \to A$ violates condition 3. This collection of pieces can easily be arranged to produce a solution, shown below.



Finally assume that only one distinct pattern appears in the pieces represented by G. If $A \to A^c$ is one of the pieces, condition 3 implies that all other pieces must be of this form. Therefore any arrangement of the pieces represents a solution and our proof is complete. □

3. Backtracking

Brandt et. al [2002] use the method of backtracking to solve Scramble Squares puzzles algorithmically. Their procedure begins by labeling the 3×3 grid with the numbers 1 through 9 in the order shown in Figure 9. The numbers stand for the order in which pieces are inserted.

The pieces are then randomly numbered 1 through 9 as well and the orientation of each piece is numbered 0 to 3 since each piece can be rotated and placed in four different ways. The first step is to place a piece into position #1 with a settled orientation. The orientation of the piece at position #1 is set to avoid repetitions obtained by rotating the whole grid.

Next, another piece is placed at position #2 with orientation 0. If the edges match, one of the remaining pieces is chosen at random for position #3 with orientation 0. This process is repeated until a piece is placed in such a way that the edges don't match. If rotating this piece 90 (or 180 or 270) degrees clockwise causes the edges to match, then the process continues. Otherwise, this piece is removed (backtracking) and a different piece is selected. If none of the pieces under any rotation makes the edges match, the previous piece is rotated 90 degrees clockwise (or removed, if its orientation number is 3) and the process continues. This trial and error process continues until all nine pieces match perfectly in their positions.

Finding the middle piece. The backtracking process described above uses randomization to select the pieces involved and thus does not take any information from the puzzle into account. We introduce a procedure called *maximizing the center* that uses information about the puzzle to potentially improve the speed of the algorithm. In the following, we will assume for simplicity of exposition that there is only one solution to a given Scramble Squares puzzle.

Notice that all of the pictures on edges in the middle of the solved puzzle will be matched and thus will need a complement, while the edges facing out on the boundary of the solved puzzle will not need a complement. Therefore we seek a procedure which will select an initial middle piece which is most likely to have matches for all four of its edge pictures.

7	8	9
6	1	2
5	4	3

Figure 9. Order of placement of the pieces in the 3 × 3 grid.

Consider a picture A and its complement A^c. Let n_A be the number of times the picture A appears on a puzzle piece and let n_{A^c} be the number of times the picture A^c appears on a puzzle piece, called the *index* of that picture. Assume without loss of generality that $n_A \geq n_{A^c}$. If x is the number of times the pattern $|A|$ appears as a complete (matched) pattern in the solution, then the probability that an occurrence of the picture A will be matched in the solution is x/n_A, while the probability that an occurrence of the picture A^c will be matched in the solution is x/n_{A^c}. Since $x/n_{A^c} \geq x/n_A$, an arbitrary occurrence of picture A^c is more likely to be matched in the solution to the puzzle than an arbitrary occurrence of picture A. Therefore, it is reasonable to select as middle position candidates pieces whose pictures have lower indices, since all four sides of the middle piece must be matched. In fact, since a picture and its complement might both have a low index, an even better measure is to use the index of the complement of a picture. This value equals the number of pictures available to be matched to the picture, and thus higher values imply more potential matches are available. The following procedure provides a method for ordering the pieces so that the ones "most likely" to be in the middle are tested there first. Of course, there are examples of puzzles in which the last piece chosen by this procedure appears in the middle, so this method is not always faster than the original backtracking method. It would be interesting to determine how frequently this method does yield some improvement over previous backtracking methods.

(1) Count the number of times each picture occurs. This is the *index* of the picture.

(2) Assign a *value index* to each piece by summing the indices of the *complements* of the pictures appearing on the piece.

(3) Place the piece with the *highest* value index in the middle.

(4) Begin with the picture on this piece whose complement has the *lowest* index.

(5) Find all the pieces containing the complement of this picture and place the one with the lowest value index next to the picture.

(6) Next use the interior picture whose complement has the lowest index from the two placed pieces and repeat step (5). Repeat the process until a picture on the interior is reached which cannot be matched to any of the remaining pieces. If no such piece exists, then the algorithm has produced a solution.

(7) If such a piece exists, rotate this piece 90 degrees clockwise and repeat. If its orientation is 3, backtrack and replace the previous piece with another piece whose value index is greater than or equal to the value index of the previous piece.

(8) Continue the procedure until arriving at a solution.

The purpose of starting with the picture whose complement has the lowest index in Steps (4) and (5) is to ensure that the picture with the highest probability of failing to find a match is tested first. Ideally this will avoid testing many extra correct pictures before finding a side of the middle piece that cannot be matched. Again, this is not a perfect strategy because it is possible for the mismatched side to be one with a high index, but perhaps this will reduce the amount of time needed to arrive at a solution for certain puzzles. Further investigation is necessary to determine the efficiency of this approach.

4. Future directions and open questions

The use of graph theory and informed backtracking to solve Scramble Squares puzzles paves the way for many new and exciting research topics. We describe several potential directions the interested reader is encouraged to explore.

Puzzles with repetitions. Repetition occurs when one picture appears two or more times in one piece. However, in a specific 2 × 2 puzzle, the solution relies on the two adjacent sides used to match other pieces. Hence, when the same picture shows up on opposite sides of a piece, while the other pictures are distinct, the solution graph properties are the same as for puzzles with no repetition. However, when the same picture appears on two adjacent sides of one piece, represented by a loop in the recording graph, different conditions are required to find a solution. While some of the conditions are similar to those for the repetition-free case, the full necessary and sufficient conditions for puzzles containing repetitions are currently unknown.

Solutions to larger puzzles. This paper focuses on solutions to 2 × 2 Scramble Squares puzzles. Certainly these results could be extended to larger puzzles in an ad hoc manner, but an ideal solution would describe conditions on a subgraph of the overall recording graph so that the subgraph corresponds to a solution.

Uniqueness. Some Scramble Squares puzzles have multiple solutions. Is it possible to find conditions under which a puzzle has a unique solution? Perhaps there is a formula using the recording graph or on the puzzle itself that enumerates the number of solutions to a given Scramble Squares puzzle. This seems to be an extremely difficult problem, but perhaps a probabilistic approach would be more likely to yield results. Such an approach would look for the probability that an arbitrary puzzle has a unique solution. Calculations could be made toward this effort by placing restrictions on the number of patterns or the number of appearances of any given pattern and then counting the number of puzzles which exhibit such properties.

It is not difficult to find conditions which are necessary for a puzzle to have at least one solution. It would be useful to have sufficient conditions as well, ideally conditions which could be easily checked using a counting argument or by verifying properties of the recording graph.

Probability. The "maximizing the center" approach will not always be faster than the depth first backtracking approach. It is possible that for some puzzles the additional information taken into account through our approach does not decrease the total time needed to solve the puzzle. If a puzzle is unusual in the sense that its central piece has the smallest value of all the pieces, then the "maximizing the center" approach would actually force us to run through all of the possible center pieces before finding the correct center piece, thus potentially taking longer than a random backtracking process. It would be very useful, therefore, to determine the probability that, given a random Scramble Squares puzzle, our approach will actually improve upon the amount of time needed to determine a solution as compared to the random backtracking approach.

References

[Brandt et al. 2002] K. Brandt, K. Burger, J. Downing, and S. Kilzer, "Using backtracking to solve the scramble squares puzzle", *J. Comput. Sci. Coll.* **17** (2002), 21–27.

[Busacker and Saaty 1965] R. G. Busacker and T. L. Saaty, *Finite graphs and networks: an introduction with applications*, McGraw-Hill, New York, 1965. MR 35 #79 Zbl 0146.20104

[Carteblanche 1947] F. de Carteblanche, "The coloured cubes problem", *Eureka* **9** (1947), 9–11.

[Grecos and Gibberd 1971] A. P. Grecos and R. W. Gibberd, "A diagrammatic solution to 'instant insanity' problem", *Math. Mag.* **44**:3 (1971), 119–124. MR 1571933 Zbl 0217.02203

[Van Deventer 1969] J. Van Deventer, "Graph theory and 'instant insanity'", pp. 283–286 in *The many facets of graph theory* (Kalamazoo, MI, 1968), edited by G. Chartrand and S. F. Kapoor, Lecture Notes in Math. **110**, Springer, Berlin, 1969. Zbl 0185.51901

Received: 2011-08-23 Revised: 2011-12-15 Accepted: 2011-12-15

masonsk@wfu.edu *Department of Mathematics, Wake Forest University, 127 Manchester Hall, Winston-Salem, NC 27109, United States*

mzmazhang@gmail.com *Davidson College, Davidson, NC 28035, United States*

INVOLVE 5:3 (2012)
dx.doi.org/10.2140/involve.2012.5.327

The n-diameter of planar sets of constant width

Zair Ibragimov and Tuan Le

(Communicated by Michael Dorff)

We study the notion of n-diameter for sets of constant width. A convex set in the plane is said to be of *constant width* if the distance between two parallel support lines is constant, independent of the direction. The Reuleaux triangles are the well-known examples of sets of constant width that are not disks. The *n-diameter* of a compact set E in the plane is

$$d_n(E) = \max\left(\prod_{1\le i<j\le n} |z_i - z_j|\right)^{\frac{2}{n(n-1)}},$$

where the maximum is taken over all $z_k \in E$, $k = 1, 2, \dots, n$. We prove that if $n = 5$, then the Reuleaux n-gons have the largest n-diameter among all sets of given constant width. The proof is based on the solution of an extremal problem for n-diameter.

1. Introduction

Sets of constant width have been an object of study by geometers for several centuries; some nontrivial examples of such sets were already known to Euler. A good summary of these studies is given in [Chakerian and Groemer 1983]; see also [Eggleston 1958]. A convex set in the plane is said to be of constant width if the distance between two parallel support lines is constant independent of the direction. Equivalently, a planar convex set W with nonempty interior is said to be of constant width if for each $\xi \in \partial W$ there exists $\eta \in \partial W$ with $|\xi - \eta| = \operatorname{diam} W$. While the disks are easily seen to be of constant width, the Reuleaux triangles are the well-known examples of sets of constant width that are not disks. In fact, sets of constant width can be thought of as generalizations of disks in that they share many properties with disks. For example, closed disks are *diametrically complete*, that is, addition of any point increases their diameter. This completeness notion characterizes the sets of constant width. Namely, the family of all complete sets is precisely the family of sets of constant width [Eggleston 1958, Theorem 52]. Another common property is

MSC2010: primary 30C65; secondary 05C25.
Keywords: n-diameter, constant width sets, Pólya extremal problem.

that sets of constant width are precisely the sets of constant diameter. (A compact set E is said to be of constant diameter if $\max\{|x-y| : y \in E\} = \operatorname{diam} E$ for each $x \in \partial E$). Also, the definition of constant width sets using parallel support lines is also based on a property of disks, namely that the distance between any two parallel support lines of a disk is constant. Finally, one more property of sets of constant width that is common with disks is that the length of the boundary arc of a disk is equal to $\lambda\pi$, where λ is the diameter of the disk; the same is true of sets of constant width. Of course, not every property of the disks is shared by sets of constant width. For example, sets of constant width $\lambda > 0$ do not have to have the same area as disks of diameter λ or that sets of constant width do not have to have smooth boundaries. In fact, by the isoperimetric inequality disks of diameter λ have the largest area while by the Blaschke–Lebesgue theorem the Reuleaux triangles of diameter λ have the smallest area. The Reuleaux triangle (named after the nineteenth-century German engineer Franz Reuleaux) of diameter λ is constructed by connecting the vertices of an equilateral triangle of sidelength λ by arcs of circles of radius λ and centered at the vertices.

Sets of constant width arise in many areas of mathematics. For instance, every odd-term Fourier series gives rise to a planar set of constant width [Kelly 1957]. Constant width sets are used in cinematography and engineering. For example, they are used in the design of the Wankel engine [Berger 1994]. They are also aesthetically pleasing, frequently turning up in art and design contexts. For example, some Irish coins have constant width shapes because of their appealing character.

The 3-diameter of sets of constant width as well as the related notions of d_3-complete sets and sets of constant 3-diameter were first studied in [Hästö et al. 2012]. As mentioned above the disks have the largest area and the Reuleaux triangles have the smallest area among all sets of given constant width. Surprisingly, the roles of the isoperimetric inequality and the Blaschke–Lebesgue theorem are reversed when it comes to 3-diameter. More precisely, among the planar sets of constant width λ, Reuleaux triangles have the largest 3-diameter, namely λ, and disks have the smallest 3-diameter, $\sqrt{3}\lambda/2$ [Hästö et al. 2012, Theorem 3.1]. On the other hand, the Reuleaux triangles have the largest area among all sets with both the diameter and 3-diameter equal to λ [Hästö et al. 2012, Proposition 2.2]. As in the case of ordinary diameter, disks are both of constant 3-diameter and d_3-complete, and d_3-complete sets are of constant 3-diameter [Hästö et al. 2012, Theorem 5.2].

In this paper we study n-diameter of sets of constant width. Our study is based on the following extremal problem: *among all planar sets of cardinality n and of diameter less than or equal to* 2*, find one with the largest n-diameter*. We conjecture that the vertices of regular n-gons have the largest n-diameter if n is odd (Conjecture 2.8) and show the conjecture is equivalent to stating that the Reuleaux n-gons have the largest n-diameter among all sets of given constant width

(Theorem 4.3). Clearly, for $n = 3$ the vertices of equilateral triangles provide a solution to the extremal problem. In contrast, the vertices of the regular 4-gon do not have the largest 4-diameter (Lemma 2.9). We show that Conjecture 2.8 holds for $n = 5$ (Theorem 3.1), and also verify the conjecture for $n = 7$ under some additional assumptions (Proposition 3.3).

2. Extremal problem for n-diameter

Definition 2.1. The n-diameter of a compact set E in the complex plane $\mathbb{C}$ is defined by

$$d_n(E) = \max\Bigg(\prod_{1\le i<j\le n} |z_i - z_j|\Bigg)^{\frac{2}{n(n-1)}},$$

where the maximum is taken over all $z_k \in E$, $k = 1, 2, \ldots, n$.

Clearly, $d_2(E)$ is the ordinary diameter of E. That is,

$$d_2(E) = \operatorname{diam} E = \sup\{|z - w| : z, w \in E\}.$$

The n-diameter is weakly decreasing in n, that is, $d_n(E) \ge d_{n+1}(E)$ [Ahlfors 1973, p. 23]; see also [Hayman 1966, Theorem 1]. We give the proof for completeness. We have

$$d_{n+1}(E) = \prod_{1\le i<j\le n+1} |z_i - z_j|^{\frac{2}{n(n+1)}};$$

thus

$$\big(d_{n+1}(E)\big)^{n(n+1)/2} = \prod_{k=2}^{n+1} |z_1 - z_k| \prod_{2\le i<j\le n+1} |z_i - z_j| \le \prod_{k=2}^{n+1} |z_1 - z_k| \,\big(d_n(E)\big)^{n(n-1)/2}.$$

Similarly, for each $l = 2, 3, \ldots, n+1$ we have

$$\big(d_{n+1}(E)\big)^{n(n+1)/2} \le \prod_{\substack{k=1\\k\ne l}}^{n+1} |z_l - z_k| \,\big(d_n(E)\big)^{n(n-1)/2}.$$

Multiplying these expressions we obtain

$$\big(d_{n+1}(E)\big)^{n(n+1)^2/2} \le \big(d_{n+1}(E)\big)^{n(n+1)}\big(d_n(E)\big)^{n(n-1)(n+1)/2}$$

which yields $d_{n+1}(E) \le d_n(E)$, as required.

The *transfinite diameter* of E is defined by

$$d_\infty(E) = \lim_{n\to\infty} d_n(E).$$

The transfinite diameter of a line segment of length L is $L/4$ and the transfinite diameter of a disk of radius r is equal to r [Ahlfors 1973, p. 28; Goluzin 1969,

p. 298]. The notion of transfinite diameter is due to Fekete and plays an important role in complex analysis. It is related to the notions of logarithmic capacity and the Chebysheff constant [Ahlfors 1973; Hille 1962; Tsuji 1959]. Some extremal problems involving the transfinite diameter and n-diameter of planar sets were studied in [Burckel et al. 2008; Dubinin 1986; Duren and Schiffer 1991; Grandcolas 2000; Grandcolas 2002; Reich and Schiffer 1964].

The simplest examples of sets for computing the n-diameter are undoubtedly the *n-tuples*, that is, sets consisting of n distinct points. Let T_n denote the set of all n-tuples in $\mathbb{C}$ of diameter less than or equal to 2.

Definition 2.2. By the extremal problem for n-diameter we mean the problem of finding

$$\sup_{E\in T_n} d_n(E).$$

According to Jung's theorem each $E \in T_n$ is contained in a disk of radius r, where $1/2 \le r \le 2/\sqrt{3}$ [Berger 1994, Theorem 11.5.8]. Also, for any $E \subset \mathbb{C}$ and for any linear transformation $L(z) = az+b$ ($a, b \in \mathbb{C}$, $a \neq 0$) we have $d_n(L(E)) = |a| d_n(E)$. Consequently,

$$\sup_{E\in T_n} d_n(E) = \sup_{E\in T'_n} d_n(E), \qquad \text{where } T'_n = \{E \in T_n : E \subset \bar{B}(0, 2/\sqrt{3})\}.$$

Since the function $d_n : T'_n \to [0, 2]$ is continuous and since T'_n is a compact subset of the n-dimensional complex space $\mathbb{C}^n$, d_n achieves its maximum in T'_n.

Definition 2.3. An n-tuple $E' \in T_n$ is called extremal if

$$\sup_{E\in T_n} d_n(E) = d_n(E').$$

Let $E_n \subset T_n$ denote the set of all extremal n-tuples in T_n. Thus, the n-diameter problem is equivalent to finding a member of E_n and computing its n-diameter.

Lemma 2.4. *Given $E \in E_n$, for each $z \in E$ there exists $w \in E$ with $|z - w| = 2$. In particular, the n-gon with vertices in E is convex.*

Proof. Let $E = \{z_1, z_2, \dots, z_n\} \in E_n$ and suppose that there exists k such that $|z_k - z_l| < 2$ for all $l = 1, 2, \dots, n$. Then there exists a disk D centered at z_k such that $|z - z_l| < 2$ for all $z \in D$ and for all $l = 1, 2, \dots, n$. Since the function

$$P(z) = \prod_{\substack{l=1\\ l\neq k}}^{n} (z - z_l)$$

is analytic in D, its modulus $|P(z)|$ cannot achieve its maximum at z_k. Hence there exists a point $z'_k \in D$ such that the n-tuple $E' = \{z_1, z_2, \dots, z'_k, \dots, z_n\}$ belongs to T_n and that $d_n(E') > d_n(E)$. Hence $E \notin E_n$, which is the required contradiction.

Let $\mathscr{C}(E)$ be the *convex hull* of E, that is, the smallest convex set containing E. Then

$$\mathscr{C}(E) = \left\{ \sum_{k=1}^{3} \lambda_k \alpha_k \mid \alpha_k \in E,\ \lambda_k \geq 0,\ \sum_{k=1}^{3} \lambda_k = 1 \right\}$$

by Carathéodory's theorem [Berger 1994, 11.1.8.6]. The first part of the lemma implies that the points $z_1, z_2, \ldots, z_n$ can only lie on the corners of $\mathscr{C}(E)$. Hence $\mathscr{C}(E)$ is the n-gon with vertices at the points $z_1, z_2, \ldots, z_n$. □

The following corollary is an immediate consequence of Lemma 2.4.

Corollary 2.5. *Let $n \geq 3$ be an odd integer. Then for each $E \in E_n$ there exist $z, w_1, w_2 \in E$ such that $|z - w_1| = |z - w_2| = 2$.*

Let $\omega = e^{2\pi i/n}$ be the nth root of unity and put

$$\mathscr{E}_n = \{1, \omega, \omega^2, \ldots, \omega^{n-1}\}.$$

Let $\overline{\mathbb{D}} = \{z \in \mathbb{C} : |z| \leq 1\}$ be the closed unit disk in $\mathbb{C}$. The following observation is credited to Pólya [Overholt and Schober 1989, p. 279]:

Theorem 2.6 (Pólya extremal problem).

$$\max d_n(\{z_1, z_2, \ldots, z_n\}) = d_n(\mathscr{E}_n) = n^{1/(n-1)}$$

where the maximum is taken over all points $z_1, z_2, \ldots, z_n$ in $\overline{\mathbb{D}}$.

Observe that

$$\operatorname{diam} \mathscr{E}_n = 2 \quad \text{if } n \text{ is even.}$$

On the other hand, if n is odd, then

$$\operatorname{diam} \mathscr{E}_n = |1 - \omega^{(n-1)/2}| = 2\sin((n-1)\pi/2n) = 2\sin(\pi/2 - \pi/2n) = 2\cos(\pi/2n).$$

Put

$$r_n = \begin{cases} 1 & \text{if } n \text{ is even,} \\ \sec(\pi/2n) & \text{if } n \text{ is odd,} \end{cases}$$

and let

$$r_n\mathscr{E}_n = \{r_n, r_n\omega, r_n\omega^2, \ldots, r_n\omega^{n-1}\}.$$

Note that

$$d_n(r_n\mathscr{E}_n) = r_n d_n(\mathscr{E}_n) = \begin{cases} n^{1/(n-1)} & \text{if } n \text{ is even,} \\ \sec(\pi/2n) n^{1/(n-1)} & \text{if } n \text{ is odd.} \end{cases} \tag{2-7}$$

Since $r_n\mathscr{E}_n \in T_n'$, we have

$$\sup_{E \in T_n} d_n(E) = \sup_{E \in T_n'} d_n(E) \geq d_n(r_n\mathscr{E}_n) = \begin{cases} n^{1/(n-1)} & \text{if } n \text{ is even,} \\ \sec(\pi/2n) n^{1/(n-1)} & \text{if } n \text{ is odd.} \end{cases}$$

Conjecture 2.8. *If n is odd, then $d_n(E) \le d_n(r_n\mathscr{E}_n)$ for each $E \in T_n$.*

Conjecture 2.8 predicts that if n is odd, then the vertices of the regular n-gons are extremal. In contrast, the vertices of the regular 4-gon are not extremal.

Lemma 2.9. *Let $E = \{r_3, r_3\omega, r_3x, r_3\omega^2\} = r_3(\mathscr{E}_3 \cup \{x\})$, where $x = 1 - \sqrt{3}$. Then $d_4(E) > d_4(\mathscr{E}_4)$.*

Proof. Recall that $r_3 = 2/\sqrt{3}$ and $\mathscr{E}_3 = \{1, \omega, \omega^2\}$, where $w = e^{2\pi i/3}$ is the third root of unity. Then $|x-1| = \sqrt{3}$ and $|x-\omega| = |x-\omega^2| = \sqrt{6-3\sqrt{3}}$. Clearly, $E \in T_4$ and hence

$$\begin{aligned} d_4(E) &= r_3(|x-1|\,|x-\omega|\,|x-\omega^2|\,|1-\omega|\,|1-\omega^2|\,|\omega-\omega^2|)^{1/6} \\ &= \frac{2}{\sqrt{3}}(|x-1|\,|x-\omega|\,|x-\omega^2|)^{1/6}\big[(|1-\omega|\,|1-\omega^2|\,|\omega-\omega^2|)^{1/3}\big]^{1/2} \\ &= \frac{2}{\sqrt{3}}(6\sqrt{3}-9)^{1/6}(d_3(\mathscr{E}_3))^{1/2} = \frac{2}{\sqrt{3}}(6\sqrt{3}-9)^{1/6}3^{1/4} \\ &= 2(2-\sqrt{3})^{1/6} > 4^{1/3} = d_4(\mathscr{E}_4), \end{aligned}$$

as required. □

Unfortunately, this idea does not seem to extend to even integers greater than 4. More precisely, some tedious computations show that if

$$E = r_{n-1}(\mathscr{E}_{n-1} \cup \{x_n\}), \qquad \text{where } x_n = 1 - 2\cos\frac{\pi}{2(n-1)},$$

then $d_n(E) < d_n(\mathscr{E}_n)$ for $n = 6, 8, 10$.

3. Cases $n = 5$ and $n = 7$

In this section we discuss Conjecture 2.8 for $n = 5$ and $n = 7$. We will make a frequent use of the well-known Ptolemy's inequality and the AM-GM inequality as well as Reinhardt's theorem. Recall that Ptolemy's inequality says that $|a-b|\,|c-d| \le |a-c|\,|b-d| + |a-d|\,|b-c|$ for all $a, b, c, d \in \mathbb{C}$ and that the equality occurs if and only if the points a, b, c, d lie on a circle in this order. The AM-GM inequality says that

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \ge (x_1 \cdot x_2 \cdots x_n)^{1/n}$$

for all nonnegative real numbers $x_1, x_2, \ldots, x_n$, and that the equality occurs when $x_1 = x_2 = \cdots = x_n$. Reinhardt's theorem says that if n is odd, then the regular n-gon has the largest perimeter among all convex n-gons of fixed diameter [Mossinghoff 2006].

Theorem 3.1. *Conjecture 2.8 is true for $n = 5$.*

Proof. Let $E = \{z_1, z_2, z_3, z_4, z_5\}$ be any 5-tuple in T_5. Without loss of generality we can assume that E is extremal, that is, $E \in E_5$. Let $\mathcal{P}_5(E)$ denote the 5-gon with vertices in E. Note that $\mathcal{P}_5(E)$ is convex by Lemma 2.4. Let P denote the perimeter of $\mathcal{P}_5(E)$. That is,

$$P = |z_1 - z_2| + |z_2 - z_3| + |z_3 - z_4| + |z_4 - z_5| + |z_5 - z_1|.$$

Clearly,

$$|z_1 - z_3|\,|z_1 - z_4|\,|z_2 - z_4|\,|z_2 - z_5|\,|z_3 - z_5| \le 2^5$$

and the equality holds if $\mathcal{P}_5(E)$ is regular. Using the AM-GM inequality we obtain

$$|z_1 - z_2|\,|z_2 - z_3|\,|z_3 - z_4|\,|z_4 - z_5|\,|z_5 - z_1| \le (P/5)^5.$$

Since $\mathcal{P}_5(E)$ is convex and $\operatorname{diam} \mathcal{P}_5(E) = 2$, by Reinhardt's theorem P is less than or equal to the perimeter of a regular 5-gon of diameter 2. Computations show that such a 5-gon has a side-length $l = \sec(\pi/5)$ and is inscribed in a circle of radius

$$r = \csc(2\pi/5) = \csc(\pi/2 - \pi/10) = \sec(\pi/10).$$

Hence

$$d_5(E) = \prod_{1 \le i < j \le 5} |z_i - z_j|^{1/10} \le \sqrt{2\sec(\pi/5)}.$$

Observe that

$$\sqrt{2\sec(\pi/5)} = \sec(\pi/10) 5^{1/4}.$$

Indeed, by Theorem 2.6 we have $d_5(\mathscr{E}_5) = 5^{1/4}$ and a direct computation yields

$$d_5(\mathscr{E}_5) = \sqrt{|1 - \omega|\,|1 - \omega^2|} = 2\sqrt{\sin(\pi/5)\sin(2\pi/5)}.$$

Hence $2\sqrt{\sin(\pi/5)\sin(2\pi/5)} = 5^{1/4}$ and it remains to show that

$$2\sec(\pi/5) = 4\sec^2(\pi/10)\sin(\pi/5)\sin(2\pi/5).$$

Equivalently,

$$\cos^2(\pi/10) = 2\cos(\pi/5)\sin(\pi/5)\sin(2\pi/5).$$

We have

$$2\cos(\pi/5)\sin(\pi/5)\sin(2\pi/5) = \sin^2(2\pi/5) = \sin^2(\pi/2 - \pi/10) = \cos^2(\pi/10),$$

as required.

Finally, the equality holds if $\mathcal{P}_5(E)$ is regular, that is, $E = L(r_5\mathscr{E}_5)$ for some $L(z) = az + b$ with $|a| = 1$. Thus,

$$d_5(E) \le \sqrt{2\sec(\pi/5)} = \sec(\pi/10)5^{1/4} = d_n(r_5\mathscr{E}_5),$$

completing the proof. □

Next, we discuss Conjecture 2.8 for $n = 7$. While we cannot verify if the conjecture is true for $n = 7$, we provide its validity under the following additional condition on 7-tuples. Given a 7-tuple $E = \{z_1, z_2, z_3, z_4, z_5, z_6, z_7\}$, suppose that the 7-gon $\mathcal{P}_7(E)$ with vertices in E is convex and that

$$\sum_{k=1}^{7} |z_k - z_{k+1}|\,|z_{k+2} - z_{k+3}| \\ \le \frac{1}{2}\sum_{k=1}^{7} |z_k - z_{k+1}|\big(|z_{k+1} - z_{k+2}| + |z_{k+3} - z_{k+4}|\big), \quad (3\text{-}2)$$

where $z_8 = z_1$, $z_9 = z_2$, $z_{10} = z_3$, $z_{11} = z_4$. Observe that the regular 7-gons satisfy condition (3-2).

Proposition 3.3. *Suppose that the 7-tuple $E = \{z_1, z_2, z_3, z_4, z_5, z_6, z_7\}$ is in T_7 and satisfies condition (3-2) and that $\mathcal{P}_7(E)$ is convex. Then*

$$d_7(E) \le 2(2\sin(\pi/14))^{1/2}(1 + 2\sin(\pi/14))^{1/6}.$$

Equality holds $\mathcal{P}_7(E)$ is a regular 7-gon of side-length

$$l = 2\sec(\pi/14)\sin(\pi/7) = 4\sin(\pi/14).$$

In particular,

$$2(2\sin(\pi/14))^{1/2}(1 + 2\sin(\pi/14))^{1/6} = \sec(\pi/14)7^{1/6}.$$

Proof. The product $\prod_{1\le k<l\le 7} |z_k - z_l|$ can be split into three parts:

$$|z_1 - z_2|\,|z_2 - z_3|\,|z_3 - z_4|\,|z_4 - z_5|\,|z_5 - z_6|\,|z_6 - z_7|\,|z_1 - z_7|,$$
$$|z_1 - z_4|\,|z_1 - z_5|\,|z_2 - z_5|\,|z_2 - z_6|\,|z_3 - z_6|\,|z_3 - z_7|\,|z_4 - z_7|,$$
$$|z_1 - z_3|\,|z_2 - z_4|\,|z_3 - z_5|\,|z_4 - z_6|\,|z_5 - z_7|\,|z_1 - z_6|\,|z_2 - z_7|.$$

It follows from the AM-GM inequality and Reinhardt's theorem that

$$|z_1 - z_2|\,|z_2 - z_3|\,|z_3 - z_4|\,|z_4 - z_5|\,|z_5 - z_6|\,|z_6 - z_7|\,|z_1 - z_7| \le (P/7)^7,$$

where P is the perimeter of a regular 7-gon with side-length $4\sin\frac{\pi}{14}$. Therefore,

$$|z_1 - z_2|\,|z_2 - z_3|\,|z_3 - z_4|\,|z_4 - z_5|\,|z_5 - z_6|\,|z_6 - z_7|\,|z_1 - z_7| \le [4\sin(\pi/14)]^7$$

and since $E \in T_7$, we also obtain

$$|z_1 - z_4|\,|z_1 - z_5|\,|z_2 - z_5|\,|z_2 - z_6|\,|z_3 - z_6|\,|z_3 - z_7|\,|z_4 - z_7| \le 2^7.$$

Moreover, the equality holds for a regular 7-gon in both of these inequalities.

It remains to find the maximum value of

$$|z_1 - z_3|\,|z_2 - z_4|\,|z_3 - z_5|\,|z_4 - z_6|\,|z_5 - z_7|\,|z_1 - z_6|\,|z_2 - z_7|.$$

To achieve this goal we will use Ptolemy's inequality. We have

$$|z_1-z_3|\,|z_2-z_4|\le|z_1-z_4|\,|z_2-z_3|+|z_1-z_2|\,|z_3-z_4|\le2|z_2-z_3|+|z_1-z_2|\,|z_3-z_4|.$$

Hence

$$|z_1-z_3|\,|z_2-z_4|\le2|z_2-z_3|+|z_1-z_2|\,|z_3-z_4|.$$

In a similar fashion we obtain

$$\begin{aligned}
|z_2-z_4|\,|z_3-z_5|&\le2|z_3-z_4|+|z_2-z_3|\,|z_4-z_5|,\\
|z_3-z_5|\,|z_4-z_6|&\le2|z_4-z_5|+|z_3-z_4|\,|z_5-z_6|,\\
|z_4-z_6|\,|z_5-z_7|&\le2|z_5-z_6|+|z_4-z_5|\,|z_6-z_7|,\\
|z_5-z_7|\,|z_1-z_6|&\le2|z_6-z_7|+|z_5-z_6|\,|z_1-z_7|,\\
|z_1-z_6|\,|z_2-z_7|&\le2|z_1-z_7|+|z_6-z_7|\,|z_1-z_2|,\\
|z_2-z_7|\,|z_1-z_3|&\le2|z_1-z_2|+|z_1-z_7|\,|z_2-z_3|.
\end{aligned}$$

Notice that the equalities hold if $\mathcal{P}_7(E)$ is a regular 7-gon in T_7, since the vertices of such a 7-gon lie on a circle and that $|z_k-z_{k+3}|=2$ for each $k=1,2,\dots,7$.

Multiplying these inequalities and applying the AM-GM inequality we obtain

$$\begin{aligned}
\prod_{1\le k\le7}|z_k-z_{k+2}|^2&=\big(|z_1-z_3|\,|z_2-z_4|\,|z_3-z_5|\,|z_4-z_6|\,|z_5-z_7|\,|z_1-z_6|\,|z_2-z_7|\big)^2\\
&\le\prod_{1\le k\le7}(2|z_{k+1}-z_{k+2}|+|z_k-z_{k+1}|\,|z_{k+2}-z_{k+3}|))\\
&\le\frac{1}{7^7}\left(2\sum_{k=1}^{7}|z_k-z_{k+1}|+\sum_{k=1}^{7}|z_k-z_{k+1}|\,|z_{k+2}-z_{k+3}|\right)^7.
\end{aligned}$$

Reinhardt's theorem implies

$$2\sum_{k=1}^{7}|z_k-z_{k+1}|\le2P=56\sin\frac{\pi}{14}$$

Once again we have the equality if $\mathcal{P}_7(E)$ is a regular 7-gon in T_7.

By our assumption we have

$$\sum_{k=1}^{7}|z_k-z_{k+1}|\,|z_{k+2}-z_{k+3}|\le\frac{1}{2}\sum_{k=1}^{7}|z_k-z_{k+1}|(|z_{k+1}-z_{k+2}|+|z_{k+3}-z_{k+4}|)$$

with equality for a regular 7-gon. Applying the AM-GM inequality for each pair of $|z_k-z_{k+1}|^2+|z_l-z_{l+1}|^2$ $(1\le k<l\le7)$ and Reinhardt's theorem we have

$$\begin{aligned}28^2\sin^2\frac{\pi}{14} &\geq \Bigl(\sum_{k=1}^{7}|z_k - z_{k+1}|\Bigr)^2\\ &\geq \frac{7}{3}\sum_{k=1}^{7}|z_k - z_{k+1}|\bigl(|z_{k+2}-z_{k+3}| + |z_{k+1}-z_{k+2}| + |z_{k+3}-z_{k+4}|\bigr)\\ &\geq 7\sum_{k=1}^{7}|z_k - z_{k+1}|\,|z_{k+2}-z_{k+3}|\end{aligned}$$

Thus,

$$\sum_{k=1}^{7}|z_k - z_{k+1}|\,|z_{k+2}-z_{k+3}| \leq 112\sin^2\frac{\pi}{14},$$

which means that

$$\prod_{k=1}^{7}|z_k - z_{k+2}| \leq \Bigl(8\sin\frac{\pi}{14} + 16\sin^2\frac{\pi}{14}\Bigr)^{7/2}.$$

This completes our proof in this case. The equality indeed occurs when z_k's are vertices of the regular 7-gon whose side-length $l = 4\sin\pi/14$. □

4. The n-diameter of sets of constant width

Let n be odd and consider the n-tuple $r_n\mathscr{E}_n$. Recall that $r_n = \sec(\pi/2n)$ and $\mathscr{E}_n = \{1, \omega, \omega^2, \dots, \omega^{n-1}\}$. Put $z_k = r_n\omega^k$, $k = 0, 1, 2, \dots, n-1$. Connect the consecutive points z_k and z_{k+1} by an arc of a circle whose center is the unique point z_l in the set $\{z_1, z_2, \dots, z_n\}$ which is equidistant from the points z_k and z_{k+1}. The radii of all such circles are the same and is equal to 2. For example, the circle centered at z_n and radius λ joins the points $z_{(n-1)/2}$ and $z_{(n+1)/2}$. The resulting set, denoted by $\mathscr{R}_n$, is of constant width 2. It is called a *Reuleaux n-gon* and the points $\{z_1, z_2, \dots, z_n\}$ are called *the vertices* of $\mathscr{R}_n$ [Chakerian and Groemer 1983, p. 59]. It follows from the construction of $\mathscr{R}_n$ that if W is any set of constant width 2 containing $r_n\mathscr{E}_n$, then $W \subset \mathscr{R}_n$. A Reuleaux n-gon of width $\lambda > 0$ is constructed in a similar fashion.

A *Reuleaux polygon* of width λ is a set of constant width λ whose boundary consists of a finitely many (necessarily odd) circular arcs of radius λ [Eggleston 1958, p. 128]. Let $\mathscr{R}$ be a Reuleaux polygon and let D be a unique disk of smallest radius containing $\mathscr{R}$. If all the corners of $\mathscr{R}$ (i.e., the intersection points of the boundary arcs of $\mathscr{R}$) are contained on ∂D, then $\mathscr{R}$ is a Reuleaux n-gon, where n is the number of corners.

Lemma 4.1. *If n is odd, then $d_n(\mathscr{R}_n) = d_n(r_n\mathscr{E}_n)$.*

Proof. Let $D = \bar{B}(0, r_n)$. Since $\mathcal{R}_n \subset D$, we have $d_n(\mathcal{R}_n) \leq d_n(D)$. Since the corners $\{z_1, z_2, \dots, z_n\}$ of $\mathcal{R}_n$ are equally spaced on ∂D, we have

$$d_n(\mathcal{R}_n) \geq \prod_{1 \leq k < l \leq n} |z_k - z_l|^{\frac{2}{n(n-1)}} = d_n(D).$$

Thus, $d_n(\mathcal{R}_n) = d_n(D) = r_n d_n(\overline{\mathbb{D}}) = \sec(\pi/2n) n^{1/(n-1)} = d_n(r_n \mathscr{E}_n)$. □

Conjecture 4.2. *If n is odd, the Reuleaux n-gons have the largest n-diameter among all sets of the same constant width.*

Theorem 4.3. *Conjecture 4.2 is equivalent to Conjecture 2.8.*

Proof. Suppose that Conjecture 2.8 is true and let W be a set of constant width λ. Let $\mathcal{R}_n(\lambda)$ be a Reuleaux n-gon of width λ. Note that if $L(z) = az + b$ with $a \neq 0$, then the set $L(W)$ is of constant width $|a|\lambda$. Let E be an n-tuple of points in ∂W with $d_n(W) = d_n(E)$. Since $(2/\lambda)E \in T_n$, using Conjecture 2.8 and Lemma 4.1 we obtain

$$d_n(W) = d_n(E) = \frac{\lambda}{2} d_n((2/\lambda)E) \leq \frac{\lambda}{2} d_n(r_n \mathscr{E}_n) = \frac{\lambda}{2} d_n(\mathcal{R}_n) = d_n(\mathcal{R}_n(\lambda)).$$

Conversely, suppose that Conjecture 4.2 is true and let E be any n-tuple in T_n. Then E is contained in a set W of constant width 2 [Eggleston 1958, Theorem 54]. Using Conjecture 4.2 and Lemma 4.1 we obtain

$$d_n(E) \leq d_n(W) \leq d_n(\mathcal{R}_n) = d_n(r_n \mathscr{E}_n). \qquad \square$$

Conjecture 4.2, if true, would imply the following corollary (see also [Hille 1962]).

Corollary 4.4. *Let $A \subset \mathbb{C}$ be any set with* $\operatorname{diam} A = \lambda > 0$ *and let D be a disk with* $\operatorname{diam} A = \lambda$. *Then*

$$d_\infty(A) \leq d_\infty(D) = \frac{\lambda}{2}.$$

Proof. Clearly, $d_\infty(D) = \lambda/2$. The set A is contained in a set W of constant width λ. If n is odd, then using (2-7), Lemma 4.1 and Conjecture 4.2 we obtain

$$d_n(A) \leq d_n(W) \leq \frac{\lambda}{2} d_n(\mathcal{R}_n(\lambda) = \frac{\lambda}{2} \sec(\pi/2n) n^{1/(n-1)}.$$

By letting n tend to infinity we obtain $d_\infty(A) \leq \lambda/2$. □

References

[Ahlfors 1973] L. V. Ahlfors, *Conformal invariants: topics in geometric function theory*, McGraw-Hill, New York, 1973. MR 50 #10211 Zbl 0272.30012

[Berger 1994] M. Berger, *Geometry, I, II*, Springer, Berlin, 1994. MR 95g:51001 Zbl 0606.51001

[Burckel et al. 2008] R. B. Burckel, D. E. Marshall, D. Minda, P. Poggi-Corradini, and T. J. Ransford, "Area, capacity and diameter versions of Schwarz's lemma", *Conform. Geom. Dyn.* **12** (2008), 133–152. MR 2010j:30050 Zbl 1233.30016

[Chakerian and Groemer 1983] G. D. Chakerian and H. Groemer, "Convex bodies of constant width", pp. 49–96 in *Convexity and its applications*, edited by P. M. Gruber and J. M. Wills, Birkhäuser, Basel, 1983. MR 85f:52001 Zbl 0518.52002

[Dubinin 1986] V. N. Dubinin, "A symmetrization method and transfinite diameter", *Sibirsk. Mat. Zh.* **27**:2 (1986), 39–46. In Russian; translated in *Sib. Math. J.* **27** (1986), 174–180. MR 88j:30053 Zbl 0595.30031

[Duren and Schiffer 1991] P. L. Duren and M. M. Schiffer, "Univalent functions which map onto regions of given transfinite diameter", *Trans. Amer. Math. Soc.* **323**:1 (1991), 413–428. MR 92d:30013 Zbl 0724.30018

[Eggleston 1958] H. G. Eggleston, *Convexity*, Cambridge Tracts in Mathematics and Mathematical Physics **47**, Cambridge University Press, New York, 1958. MR 23 #A2123 Zbl 0086.15302

[Goluzin 1969] G. M. Goluzin, *Geometric theory of functions of a complex variable*, Translations of Mathematical Monographs **26**, American Mathematical Society, Providence, RI, 1969. MR 40 #308 Zbl 0183.07502

[Grandcolas 2000] M. Grandcolas, "Regular polygons and transfinite diameter", *Bull. Austral. Math. Soc.* **62**:1 (2000), 67–74. MR 2001k:52005 Zbl 0976.52004

[Grandcolas 2002] M. Grandcolas, *Problems of diameters in the plane, the Cauchy problem for a derivor*, thesis, Université de Rouen, Mont-Saint Aignan, 2002.

[Hästö et al. 2012] P. Hästö, Z. Ibragimov, and D. Minda, "Convex sets of constant width and 3-diameter", *Houston J. Math* **38**:2 (2012), 421–443.

[Hayman 1966] W. K. Hayman, *Transfinite diameter and its applications*, Matscience Report **45**, Institute of Mathematical Sciences, Madras, 1966. MR 52 #14272

[Hille 1962] E. Hille, *Analytic function theory*, vol. 2, Ginn, Boston, 1962. MR 34 #1490 Zbl 0102.29401

[Kelly 1957] P. J. Kelly, "Curves with a kind of constant width", *Amer. Math. Monthly* **64** (1957), 333–336. MR 19,1073b Zbl 0080.15602

[Mossinghoff 2006] M. J. Mossinghoff, "A $1 problem", *Amer. Math. Monthly* **113**:5 (2006), 385–402. MR 2006m:51021 Zbl 1170.51007

[Overholt and Schober 1989] M. Overholt and G. Schober, "Transfinite extent", *Ann. Acad. Sci. Fenn. Ser. A I Math.* **14**:2 (1989), 277–290. MR 90m:30028 Zbl 0699.30009

[Reich and Schiffer 1964] E. Reich and M. Schiffer, "Estimates for the transfinite diameter of a continuum", *Math. Z.* **85** (1964), 91–106. MR 30 #4921 Zbl 0129.29304

[Tsuji 1959] M. Tsuji, *Potential theory in modern function theory*, Maruzen, Tokyo, 1959. MR 22 #5712 Zbl 0087.28401

Received: 2011-09-12 Revised: 2011-12-14 Accepted: 2011-12-16

zibragimov@fullerton.edu
Department of Mathematics, California State University, Fullerton, McCarthy Hall 154, Fullerton, CA 92831, United States

tuanl@wpi.edu
Department of Mathematical Sciences,
Worcester Polytechnic Institute, 100 Institute Road,
Worcester, MA 01609, United States

INVOLVE 5:3 (2012)
dx.doi.org/10.2140/involve.2012.5.339

Boolean elements in the Bruhat order on twisted involutions

Delong Meng

(Communicated by Kenneth S. Berenhaut)

We prove that a permutation in the Bruhat order on twisted involutions is Boolean if and only if it avoids the following patterns: 4321, 3421, 4312, 4231, 32541, 52143, 351624, 456123, 426153, 321654, 561234, 345612, 3416275, 3561274, 1532746, 4517236, 34127856, 35172846, and 36712845. This result answers a question proposed by Hultman and Vorwerk. Our technique provides an application of the pictorial representation of the Bruhat order given by Incitti.

1. Introduction

In this paper, we answer the following question proposed by Hultman and Vorwerk [2009, Problem 5.1].

Problem. A permutation $w \in \mathfrak{S}_n$ is said to be a *twisted involution* if ww_0 is an involution, where $w_0 = n, n-1, \dots, 1$. Let $Tw(\mathfrak{S}_n)$ denote the Bruhat order on twisted involutions. With pattern avoidance, classify all twisted involutions whose principal order ideal in $Tw(\mathfrak{S}_n)$ is Boolean.

We first define some requisite terms in the problem statement (see [Björner and Brenti 2005] for background reading).

Definition. Let

$$l(w) = |\{\,\{i, j\} : i < j \text{ and } w(i) > w(j)\}|$$

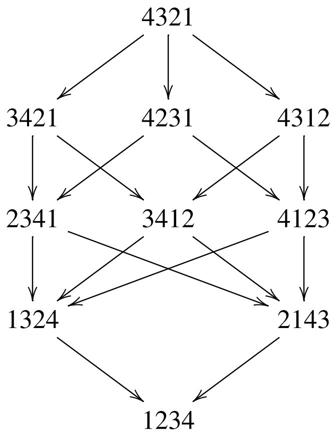
denote the number of inversions of w. The *Bruhat order of the symmetric group*, denoted by $Br(\mathfrak{S}_n)$, is a partial order on $\mathfrak{S}_n$ defined as follows: w covers w' if and only if $l(w) = l(w') + 1$ and w is obtained from w' by a transposition of $w'(i)$ and $w'(j)$ for some $1 \le i, j \le n$.

MSC2010: 05E15.
Keywords: pattern avoidance, Bruhat order, twisted involutions, Boolean posets.

Definition. The *Bruhat order on twist involutions*, denoted by $Tw(\mathfrak{S}_n)$, is the poset on twisted involutions defined by $u \leq v$ in $Tw(\mathfrak{S}_n)$ if and only if $u \leq v$ in $Br(\mathfrak{S}_n)$.

Let Q be a poset. The *principal order ideal* of $w \in Q$, denoted by $P_Q(w)$ (or $P(w)$ when the context is clear), is the subposet of Q induced by the set of elements less than or equal to w. The Boolean poset B_k is the poset on the subsets of $\{1, 2, \dots, k\}$ partially ordered by inclusion. A twisted involution w is said to be *Boolean* if its principal order ideal in $Tw(\mathfrak{S}_n)$ is isomorphic to a Boolean poset.

Example. The Boolean elements of $Tw(\mathfrak{S}_4)$ are 2341, 3412, 4123, 1324, 2143, and 1234.



We now present a brief history of this problem. Tenner [2007] brought pattern avoidance to the study of the Bruhat order. She classified Boolean elements of $Br(\mathfrak{S}_n)$ as 321- and 3412-avoiding permutations. Hultman and Vorwerk [2009] studied an analogue of Tenner's result for involutions: 4321-, 45312-, and 456123-avoiding permutations. At the end of [Hultman and Vorwerk 2009], the authors asked whether a similar result exists for twisted involutions. We settle this question with the following theorem.

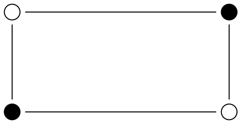
Theorem 1.1. *A twisted involution is Boolean if and only if it avoids all forbidden patterns. The forbidden patterns are* 4321, 3421, 4312, 4231, 32541, 52143, 351624, 456123, 426153, 321654, 561234, 345612, 3416275, 3561274, 1532746, 4517236, 34127856, 35172846, *and* 36712845.

As a side note, the poset $Tw(\mathfrak{S}_n)$ is isomorphic to the dual of $I(\mathfrak{S}_n)$, the Bruhat order on involutions. Consequently, our result also characterizes Boolean principal order filters of $I(\mathfrak{S}_n)$.

Previous work [Hultman and Vorwerk 2009; Tenner 2007] relied heavily on the algebraic properties of Coxeter groups. We prove Theorem 1.1 using permutation diagrams that represent the cover relations in the Bruhat order. Such diagrams were introduced by Incitti [2003; 2004; 2005].

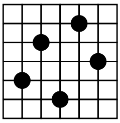
Let $w \in \mathfrak{S}_n$. The *permutation diagram* of w is the set of lattice points $(i, w(i))$, where $1 \leq i \leq n$. Permutation diagrams of twisted involutions of $\mathfrak{S}_n$ are symmetric about $x + y = n + 1$.

Incitti [2005] shows that w covers w' in $Br(\mathfrak{S}_n)$ if and only if their permutation diagrams differ by the following rectangle.

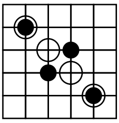


White dots belong to w and black to w', and no points lie inside the rectangle. Call the above rectangle a *cover block*.

Example. Left is the permutation diagram of 24153. The picture to the right shows that 4321 covers 4231 in $Br(\mathfrak{S}_4)$.



24153



4321>4231

○ > ●

Incitti [2004] classifies the cover relations of $I(\mathfrak{S}_n)$ with six types of cover blocks. Reflecting them about the line $x = (n+1)/2$ gives us the cover blocks of $Tw(\mathfrak{S}_n)$ (see Section 3).

The key idea of our proof of Theorem 1.1 is to examine pairs of cover blocks on the same permutation diagram. To illustrate our technique, we start with an alternative and arguably simpler proof of Tenner [2007, Theorem 5.3] (see Section 2). In particular, we remove the need for Tenner's characterization of vexillary permutations [Tenner 2006, Theorem 3.8].

Section 3 contains the proof of Theorem 1.1. We discuss further directions in Section 4.

The Bruhat order on Coxeter groups is an extensively studied subject in combinatorics (see [Björner and Brenti 2005]). Following the work of Richardson and Springer [1990], there has been a surge of interest in the Bruhat order on twisted involutions because of its application to algebraic geometry and its resemblance to the Bruhat order on the symmetric group. For example, Hultman [2005; 2007] showed that $Tw(\mathfrak{S}_n)$ satisfies the deletion property and the subword property known for $Br(\mathfrak{S}_n)$. The similarity between $Tw(\mathfrak{S}_n)$ and $Br(\mathfrak{S}_n)$ inspired Hultman and Vorwerk [2009] to propose this problem.

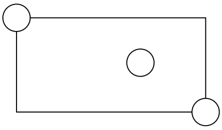
2. Boolean elements of the symmetric group

In this section, we give an alternative proof of Tenner [2007, Theorem 5.3] to illustrate our approach to Theorem 1.1.

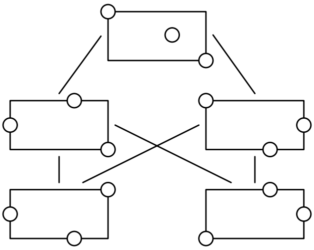
Definition. Elements of a poset are called *vertices* and cover relations *edges*. An edge uv is called an *upward move from* u if v covers u. Define a *downward move* similarly. Two edges are said to *commute uniquely* if they lie on a unique 4-cycle.

Theorem [Tenner 2007, Theorem 5.3]. *A permutation $w \in Br(\mathfrak{S}_n)$ is Boolean if and only if w is* 321- *and* 3412-*avoiding.*

Proof of necessity. Suppose w contains a 321- or 3412-pattern. Let u denote the minimal element of $P(w)$ that contains a 321- or 3412-pattern. Then the permutation diagram of u contains the following figure, where no other points lie inside the rectangle.



The two downward edges from u that act on this rectangle do not commute uniquely, as shown below.



Therefore, w is not Boolean. □

Proof of sufficiency. We start with the following lemma.

Lemma 2.1. *Let P be a graded and connected poset. If every pair of edges that share a vertex in P commute uniquely, then P is a Boolean poset.*

Proof. Let M denote the maximal element of P (the maximum exists because upward moves commute uniquely). Suppose M covers $m_1, m_2, \dots, m_k$. Define $f(u) = \{i : m_i \geq u\}$. We prove that f bijectively maps the i-th row of P to the i-th row of B_k using strong induction on i. Base case is trivial.

Suppose f bijectively maps the first i rows of P to the first i rows of B_k. Let u be an element in the $(i+1)$st row. We claim that an element v covers u if and only if $f(v) = f(u) \backslash x$ for some $x \in f(u)$.

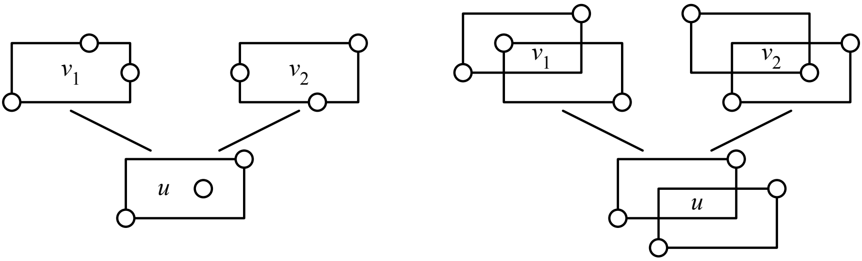
If u is covered with $v_1, v_2, \dots, v_j$, then $f(u) = \bigcup_{1 \le a \le j} f(v_a)$. Since for all $1 \le a, b \le j$, the upward moves uv_a and uv_b commute uniquely, we have $f(v_a)$ and $f(v_b)$ differ by exactly one element. Therefore, for all $1 \le a \le j$, we have $f(v_a) = f(u) \backslash x$ for some $x \in f(u)$. Conversely, suppose $f(v) = f(u) \backslash x$ for some $x \in f(u)$. Let

$$v' := f^{-1}(f(v_j) \cap f(v)).$$

Since $v_j v'$ and $v_j u$ commute uniquely, there exists a v'' in the i-th row such that $v'v''uv_j$ is a four cycle. Then $f(v'') = f(u) \backslash x$. Since the i-th row of P is isomorphic to the i-th row of B_k, we have $v = v''$.

Therefore, f maps the $(i+1)$st row of P to the $(i+1)$st row of B_k, and P is isomorphic to the Boolean post B_k. □

Suppose w is 321- and 3412-avoiding. It is easy to check that all $u \le w$ are 321- and 3412-avoiding. If u is 321- and 3412-avoiding, then all pairs of downward moves from u commute uniquely. An upward move from u commute uniquely with a downward move from u. Suppose two upward moves uv_1 and uv_2 do not commute uniquely, then these two moves must be applied to one of the following figures.



In the right diagram, the element greater than both v_1 and v_2 must contain a 321-pattern, so only one of v_1 or v_2 belongs to $P(w)$.

In the left diagram, there exist v_1' and v_2' that cover both v_1 and v_2. Let v be the element that covers v_1' and v_2'. Then vv_1' and vv_2' are two downward edges that do not commute uniquely, as shown below.

$$\begin{tikzcd}
 & v \arrow[dl] \arrow[dr] & \\
v_1' \arrow[d] \arrow[drr] & & v_2' \arrow[d] \arrow[dll] \\
v_1 \arrow[dr] & & v_2 \arrow[dl] \\
 & u &
\end{tikzcd}$$

Thus, one of v_1' and v_2' does not belong to $P(w)$, and uv_1 and uv_2 do commute uniquely in $P(w)$.

Therefore, all pairs of edges commute uniquely in $P(w)$, and w is Boolean by Lemma 2.1. □

Remark. The key idea of the proof of necessity is to identify a pair of downward moves that do not commute uniquely. The proof of sufficiency follows from Lemma 2.1.

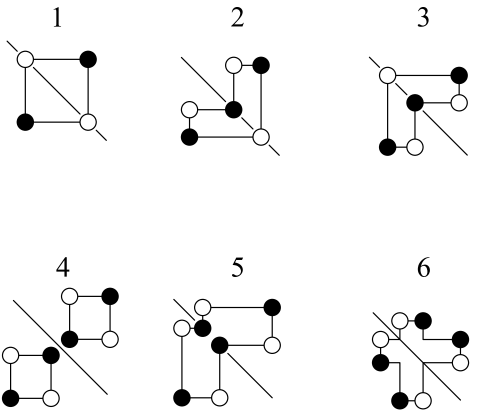
We can also use this idea to prove Hultman and Vorwerk [2009, Theorem 1.1], with the caveat that there are six types of cover blocks in the Bruhat order on involutions.

3. Proof of the main theorem

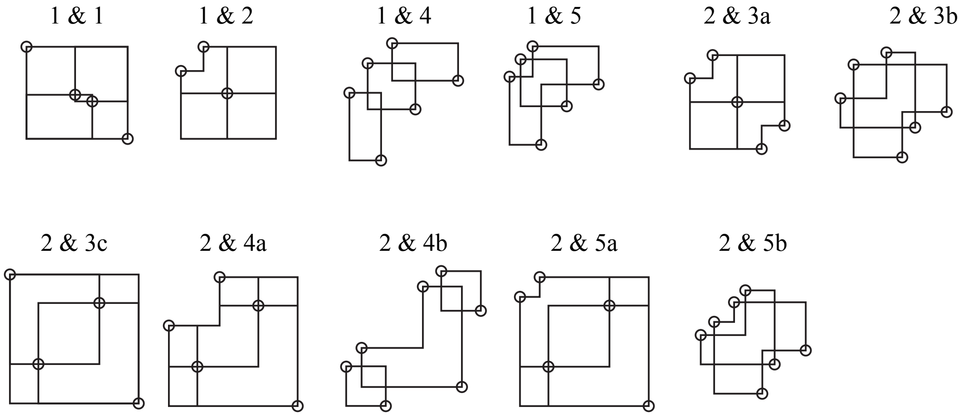
We wish to apply the same technique as the previous section, so we need to identify pairs of edges that do not commute uniquely.

We first classify the cover blocks of the Bruhat order on twisted involutions. Incitti [2004] characterizes the six types of cover blocks of the Bruhat order on involutions. Since permutation diagrams of twisted involutions are equivalent to those of involutions reflected about $x = (n+1)/2$, we obtain the following characterization of cover relations of $Tw(\mathfrak{S}_n)$.

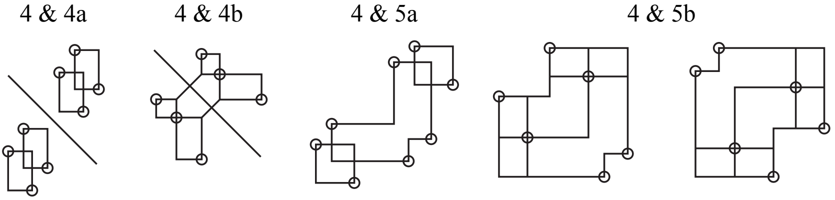
Definition. Let w and w' denote two twisted involutions. We have w covers w' if and only if their permutation diagrams differ by one of the following *cover blocks*. The white dots belong to w and black to w', and no points lie inside these shapes.

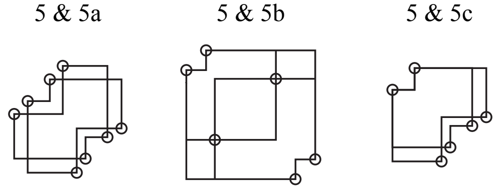


The six types of cover blocks induce fifteen types of intersecting pairs. The pairs of downward moves that do not commute uniquely define the *forbidden pairs* shown below, where no points lie inside the area enclosed by the lines.



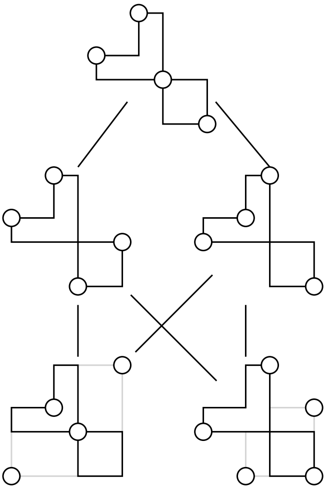
3 & 1, 3 & 4, 3 & 5 are rotations of 2 & 1, 2 & 4, 2 & 5 by 180 degrees





These forbidden pairs will give us the forbidden patterns in Theorem 1.1. Note that cover relation 6 never appears in the forbidden pairs because any forbidden pair with relation 6 induces a forbidden pattern already contained in the earlier ones.

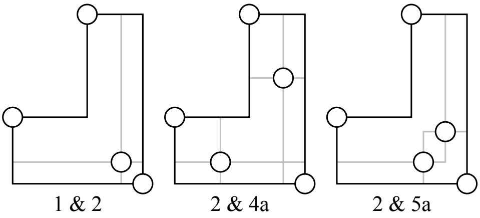
Example. Two downward moves applied to the forbidden pair 1&2 do not commute uniquely. (See figure on the right.)



We now prove Theorem 1.1.

Proof of necessity. Suppose w contains one of the forbidden patterns. If this forbidden pattern in the permutation diagram of w is not symmetric about $x+y=n+1$, then we can treat the corresponding points as the full Bruhat order. Since all forbidden patterns contain either 321 or 3412, the permutation cannot be Boolean.

We now assume that the forbidden pattern is symmetric about $x+y=n+1$. Let u denote the minimal element of $P(w)$ that contains a forbidden pattern. Then the permutation diagram of u contains a forbidden pair. (The forbidden pattern of u with the smallest area in the permutation diagram cannot contain other points. For example, if there are other points inside cover block 2, then we obtain a forbidden pattern with smaller area, as shown below.)



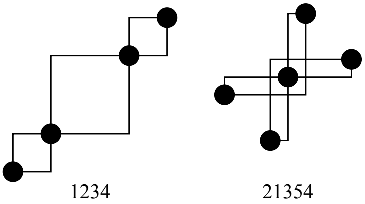
The two downward moves applied to this forbidden pair of u do not commute uniquely. Thus, w is not Boolean. □

Proof of sufficiency. To apply Lemma 2.1, we check two things:

1. If w avoids all forbidden patterns, then any $u \leq w$ also avoids all forbidden patterns.
2. If u avoids all forbidden patterns, then all pairs of edges emanating from u commute uniquely.

Any $u \leq w$ avoids all forbidden patterns because, after every downward move, the number of cover blocks and the rank of u decrease by exactly one, which means no new cover block can be created.

If u avoids all forbidden patterns, then two downward moves from u commute uniquely. We can check that an upward move and a downward move always commute (this also follows from the lexicographical shellability of $Tw(\mathfrak{S}_n)$). The pairs of upward moves that do not commute uniquely induce patterns 1234, 21354, and 321654.



We have 321654 is a forbidden pattern. For the other two patterns, the element that covers both end points of an upward move contains a element with a forbidden pattern. Therefore, only one of the endpoints is contained in $P(w)$, and all pairs of upward moves do commute uniquely in $P(w)$. Lemma 2.1 shows that w is indeed Boolean. □

4. Further remarks

Our result provides an application of Incitti's pictorial representation of the Bruhat order [Incitti 2004]. Incitti [2003; 2004; 2005] also classify representations of cover relations for the Bruhat order on Coxeter groups of types B and D as well as involutions in these groups.

Question. What is the analogue of our result for Coxeter groups of types B and D?

Green and Losonczy [2002] classify Boolean elements of the poset on commutation classes of reduced decompositions: 4321-, 4231-, 4312-, and 3421-avoiding permutations. The author's recent work [Meng ≥ 2012] generalizes Green and Losonczy's work to the higher Bruhat order. Using Incitti's pictures, we can show that $Br(\mathfrak{S}_n)$ is generated by 4-cycles. Compare this with the fact that the higher Bruhat order $B(n, 2)$ is generated by 4-cycles and 8-cycles, we believe that studying the similarity between the strong Bruhat order and the higher Bruhat order is worthwhile.

Acknowledgements

This research was carried out at the University of Minnesota Duluth under the supervision of Joseph Gallian, with the financial support of the National Science Foundation and the Department of Defense (grant number DMS 0754106), the National Security Agency (grant number H98230-06-1-0013), and the MIT Department of Mathematics. The author thanks Joseph Gallian for his encouragement and support, and Mike Develin, Nathan Kaplan, Nathan Pflueger, and Alex Zhai for helpful discussions.

References

[Björner and Brenti 2005] A. Björner and F. Brenti, *Combinatorics of Coxeter groups*, Graduate Texts in Mathematics **231**, Springer, New York, 2005. MR 2006d:05001 Zbl 1110.05001

[Green and Losonczy 2002] R. M. Green and J. Losonczy, "Freely braided elements of Coxeter groups", *Ann. Comb.* **6**:3-4 (2002), 337–348. MR 2004d:20042 Zbl 1052.20028

[Hultman 2005] A. Hultman, "Fixed points of involutive automorphisms of the Bruhat order", *Adv. Math.* **195**:1 (2005), 283–296. MR 2006a:06001 Zbl 1102.06002

[Hultman 2007] A. Hultman, "The combinatorics of twisted involutions in Coxeter groups", *Trans. Amer. Math. Soc.* **359**:6 (2007), 2787–2798. MR 2007k:20082 Zbl 1166.20030

[Hultman and Vorwerk 2009] A. Hultman and K. Vorwerk, "Pattern avoidance and Boolean elements in the Bruhat order on involutions", *J. Algebraic Combin.* **30**:1 (2009), 87–102. MR 2011c:06009 Zbl 1225.06002

[Incitti 2003] F. Incitti, "The Bruhat order on the involutions of the hyperoctahedral group", *European J. Combin.* **24**:7 (2003), 825–848. MR 2004h:05128 Zbl 1056.20027

[Incitti 2004] F. Incitti, "The Bruhat order on the involutions of the symmetric group", *J. Algebraic Combin.* **20**:3 (2004), 243–261. MR 2005h:06003 Zbl 1057.05079

[Incitti 2005] F. Incitti, "Bruhat order on classical Weyl groups: minimal chains and covering relation", *European J. Combin.* **26**:5 (2005), 729–753. MR 2006b:06005 Zbl 1083.20036

[Meng $\geq$ 2012] D. Meng, "Reduced decompositions and permutation patterns generalized to the higher Bruhat order", http://web.mit.edu/delong13/papers.html. Submitted.

[Richardson and Springer 1990] R. W. Richardson and T. A. Springer, "The Bruhat order on symmetric varieties", *Geom. Dedicata* **35**:1-3 (1990), 389–436. MR 92e:20032 Zbl 0704.20039

[Tenner 2006] B. E. Tenner, "Reduced decompositions and permutation patterns", *J. Algebraic Combin.* **24**:3 (2006), 263–284. MR 2007f:05008 Zbl 1101.05003

[Tenner 2007] B. E. Tenner, "Pattern avoidance and the Bruhat order", *J. Combin. Theory Ser. A* **114**:5 (2007), 888–905. MR 2008d:05164 Zbl 1146.05054

Received: 2011-09-17 Accepted: 2011-09-22

delong13@mit.edu

*Department of Mathematics, Massachusetts Institute
of Technology, 77 Massachusetts Avenue,
Cambridge, MA 02139, United States*

INVOLVE 5:3 (2012)
dx.doi.org/10.2140/involve.2012.5.349

Statistical analysis of diagnostic accuracy with applications to cricket

Lauren Mondin, Courtney Weber, Scott Clark,
Jessica Winborn, Melinda M. Holt and Ananda B. W. Manage

(Communicated by Scott Chapman)

In the sport of cricket, as with any other sport, spectators and officials would like the games to be as fair as possible. To this end, we evaluate methods used to determine the winner of interrupted games using statistical accuracy. In the traditional One Day International cricket matches, the current Duckworth–Lewis (DL) method and the discounted most productive overs (DMPO) method are each used for predicting the winner. However, with the growing popularity of shorter Twenty20 matches, a new Bhattacharya–Gill–Swartz (BGS) method has also been introduced. We created both classical and Bayesian intervals to estimate the true accuracy of each. Using past game data from 2005–2010, we compared the DL, DMPO and BGS methods using the new accuracy intervals and receiver operating characteristic (ROC) curves.

1. Introduction

This paper examines the accuracy of methods for predicting the winner of interrupted cricket matches. Cricket is a field game that was first seen in 16th century England and, within 200 years, became England's national sport. International games were being held by the 19th century and the popularity has only grown since then, making it now the second most popular sport in the world, next to association football. The International Cricket Council (ICC) is the governing body, currently with 104 members. It was formed in 1909 by England, Australia, and South Africa, who originally called it the Imperial Cricket Conference.

There are many variations of cricket; however, Test cricket, One Day International (ODI), and Twenty20 are the most common. Test cricket is the longest form of cricket and can last up to five days. One Day International is the most common, lasting roughly eight hours with a maximum of 50 overs in each of two innings. The

MSC2010: 62PXX.
Keywords: interval estimates, Bayesian methods, ROC curves, delta method.

newest form of cricket, Twenty20, has two innings with a maximum of 20 overs each, lasting only about 4 hours and growing in popularity.

Similarities between cricket and baseball can be found, even though the two games differ in many ways. Both sports have batters and pitchers, although in cricket the pitchers are called bowlers. A bowler does not pitch the ball like a baseball pitcher would; he bowls it with a stiff arm. In baseball, there are nine innings with the teams alternating offense and defense in each inning. In cricket, the word "innings" is both singular and plural, with most variations of cricket consisting of only two innings. Additionally, only one team bats per innings. Therefore, the second team's goal is to beat the first team's score before they lose 10 wickets or have been bowled all of their overs. Six bowls are equivalent to one over, so, for example, in ODI there are 50 overs in an innings or 300 bowls. Cricket also differs from baseball in that there are 11 players on a team, and a coin toss at the beginning determines who bats first. Lastly, the batsmen bat in pairs and score by running back and forth between the wickets or by hitting the ball outside the boundary. The 10 outs per innings are obtained primarily by knocking over the wickets.

As with any sport played outside, games can be interrupted due to rain or other bad weather. When this happens, a winner must be decided. However, it would be unfair to simply decide the winner based on current points, since one of the teams may have had a full innings to score and the other has not had this opportunity. There are several methods that can be used to declare a winner. The most popular is the Duckworth–Lewis method, which is used for 50-over games and can be scaled down for 20-over games. A new method for 20-over games is the Bhattacharya–Gill–Swartz method, which, while based on the Duckworth–Lewis method, is designed specifically for 20-over games instead of 50-over games. A third method, used for 50-over games, is the discounted most productive overs (DMPO) method, which tends to favor the first team batting. We hope to see an improvement for 20-over cricket games. These three methods are compared to determine the most accurate method for determining the winner of Twenty20 cricket games.

2. Target methods for interrupted matches

Due to the nature of cricket, interrupted matches often result from inclement weather. When a game has been stopped, and thus shortened, the team batting does not get their intended number of overs. Therefore, the winner of the game must be predicted by calculating the second team's target score, which takes into account their unused overs. Since cricket is such a historic and popular game, several methods of predicting the winner have been developed.

The most popular method of choosing a winner in an interrupted game is the Duckworth–Lewis (DL) method [1998]. This method is currently the preferred

	Wickets Lost				
Overs Left	0	2	5	7	9
50	100	85.1	49	22	4.7
40	89.3	77.8	47.6	22	4.7
30	75.1	67.3	44.7	21.8	4.7
20	56.6	52.4	38.6	21.2	4.7
10	32.1	30.8	26.1	17.9	4.7

Table 1. DL resource table for ODI cricket.

method of the ICC and is based on the number of overs remaining in the game (u) and the amount of wickets that have been lost (w). This relationship follows an exponential decay model $Z(u, w) = Z_0(w)\big(1 - \exp\{-b(w)u\}\big)$, where $b(w)$ is the exponential decay constant and $Z_0(w)$ is the asymptotic average from the wickets remaining in unlimited overs. Table 1 is a shortened version of the DL table used for ODI cricket. To calculate the target score of a game that was interrupted after 10 overs were bowled and 5 wickets were lost, find the row indicating that there are 40 overs left and the column indicating there have been 5 wickets lost. The common entry cell provides the percentage of resources remaining, so the target score in this situation would be the first team's final score multiplied by $(1-0.476)$. If this calculated number, or "target score," is greater than the second team's score at the time the game was interrupted, the second team loses; otherwise, the second team wins. The original DL table is based on a game that consists of innings with 50 overs. Here we consider games with innings of 20 overs. Currently, the ICC rescales the table from 50 overs to allow for matches with 20 overs in an innings. Table 1 is an example of a partial DL table for matches with 50 overs.

The DL table for Twenty20 cricket is presented in Table 2.

The Bhattacharya–Gill–Swartz (BGS) method [Bhattacharya et al. 2010] is a new method that is similar to the DL method, but developed specifically for Twenty20 matches. This method calculates r_{uw}, the estimated percentage of resources remaining when u overs are available and w wickets have been taken. To impose necessary monotonicity constraints on rows and columns, BGS employs isotonic regression, minimizing $F = \sum\sum q_{uw}(r_{uw} - y_{uw})^2$ with respect to the matrix $Y = (y_{uw})$, for $u = 1, \dots, 20$ and $w = 0, \dots, 9$, where q_{uw} are weights based on sample variances. The resulting matrix, Y, gives an unsatisfactory table of values, as with the Duckworth–Lewis table, with some adjacent entries being the same. Through Gibbs sampling and a Bayesian model, a new table consisting of the estimated posterior means is offered by [Bhattacharya et al. 2010] as an alternative to the Duckworth–Lewis table for Twenty20 cricket. Due to the lack of games available, some entries in the BGS table were initially left empty, then later filled

	Wickets Lost									
Overs Left	0	1	2	3	4	5	6	7	8	9
20	100	96.8	92.6	86.7	78.8	68.2	54.4	37.5	21.3	8.3
19	96.1	93.3	89.2	83.9	76.7	66.6	53.5	37.3	21.0	8.3
18	92.2	89.6	85.9	81.1	74.2	65.0	52.7	36.9	21.0	8.3
17	88.2	85.7	82.5	77.9	71.7	63.3	51.6	36.6	21.0	8.3
16	84.1	81.8	79.0	74.7	69.1	61.3	50.4	36.2	20.8	8.3
15	79.9	77.9	75.3	71.6	66.4	59.2	49.1	35.7	20.8	8.3
14	75.4	73.7	71.4	68.0	63.4	56.9	47.7	35.2	20.8	8.3
13	71.0	69.4	67.3	64.5	60.4	54.4	46.1	34.5	20.7	8.3
12	66.4	65.0	63.3	60.6	57.1	51.9	44.3	33.6	20.5	8.3
11	61.7	60.4	59.0	56.7	53.7	49.1	42.4	32.7	20.3	8.3
10	56.7	55.8	54.4	52.7	50.0	46.1	40.3	31.6	20.1	8.3
9	51.8	51.1	49.8	48.4	46.1	42.8	37.8	30.2	19.8	8.3
8	46.6	45.9	45.1	43.8	42.0	39.4	35.2	28.6	19.3	8.3
7	41.3	40.8	40.1	39.2	37.8	35.5	32.2	26.9	18.6	8.3
6	35.9	35.5	35.0	34.3	33.2	31.4	29.0	24.6	17.8	8.1
5	30.4	30.0	29.7	29.2	28.4	27.2	25.3	22.1	16.6	8.1
4	24.6	24.4	24.2	23.9	23.3	22.4	21.2	18.9	14.8	8.0
3	18.7	18.6	18.4	18.2	18.0	17.5	16.8	15.4	12.7	7.4
2	12.7	12.5	12.5	12.4	12.4	12.0	11.7	11.0	9.7	6.5
1	6.4	6.4	6.4	6.4	6.4	6.2	6.2	6.0	5.7	4.4

Table 2. DL resource table for Twenty20 cricket.

with the corresponding DL entries. The resulting BGS table is presented in Table 3.

A third, less commonly applied, method offered for comparison is the discounted most productive overs (DMPO) method. This method offers a new way of calculating a target score for the second team after the game is interrupted. First, order the overs according to the number of runs per over, highest to lowest. The target score for Team 2 is then found by summing the same number of highest scoring overs of Team 1 and then discounted by 0.5% per over lost. This method tends to favor Team 1 for 50-over cricket games. The DMPO method will be considered herein to determine whether or not its performance is competitive with the DL and BGS methods for 20-over cricket games.

3. Large-sample confidence intervals for accuracy

To compare the performance of the DL, BGS and DMPO methods when predicting winners in 20-over cricket, we compared their predictions to actual final results for 120 recent uninterrupted Twenty20 matches from 2005–2010, obtained from

	Wickets Lost									
Overs Left	0	1	2	3	4	5	6	7	8	9
20	100	96.9	93.0	87.9	81.3	72.2	59.9	44.8	29.7	17.6
19	95.6	90.9	87.7	83.0	76.9	68.3	56.5	42.0	27.2	15.3
18	91.7	86.7	82.9	78.7	73.2	65.4	54.2	40.2	25.7	13.9
17	87.7	82.3	78.9	73.8	69.7	62.8	52.2	38.7	24.6	12.8
16	83.5	78.2	75.3	70.5	66.4	60.2	50.3	37.4	23.5	12.0
15	79.2	74.3	70.9	66.9	62.6	57.4	48.4	36.2	22.7	11.2
14	75.1	70.7	67.3	63.7	59.3	54.6	46.4	35.0	21.8	10.5
13	71.5	67.4	63.6	60.3	56.2	51.5	44.3	33.8	21.0	9.8
12	68.3	63.7	60.2	56.8	52.9	47.5	41.9	32.6	20.2	9.1
11	65.0	59.9	56.6	53.3	49.7	43.9	39.3	31.3	19.4	8.5
10	61.3	56.0	52.6	50.1	46.0	40.8	36.1	30.0	18.6	7.9
9	57.9	52.3	47.9	46.1	42.5	37.8	33.1	28.3	17.7	7.2
8	54.0	48.3	44.3	41.7	38.9	34.9	30.2	26.1	16.7	6.6
7	49.3	44.2	40.2	37.4	35.4	32.1	27.2	23.4	15.7	5.9
6	41.7	38.5	35.7	33.0	31.7	29.0	24.2	20.0	14.5	5.2
5	36.2	33.4	31.0	28.6	27.3	25.5	21.5	17.0	12.2	4.4
4	30.8	28.0	26.1	24.1	22.4	20.7	18.3	14.2	10.0	4.4
3	25.4	22.8	21.1	19.4	17.7	16.5	14.4	11.6	7.9	2.5
2	19.7	17.2	15.5	14.1	12.7	11.9	10.6	9.3	6.2	1.6
1	13.7	11.3	9.7	8.5	7.3	6.7	6.0	5.2	4.2	0.9

Table 3. BGS resource table for Twenty20 cricket.

www.cricket.org. While the DL, BGS and DMPO methods are sometimes used for cricket matches interrupted more than once before being stopped, we considered their performance for cricket matches that were hypothetically interrupted only once during the second team's bat. For comparison purposes, we considered the predicted outcomes of each game for three possible situations: interruption at 5 overs, 10 overs and 15 overs. Because the effect of team superiority differences should be kept at a minimum, only 10 countries were included in the sample of 120 cricket matches. These countries are Australia, Bangladesh, Pakistan, South Africa, West Indies, Sri Lanka, India, New Zealand, England and Zimbabwe.

We compared the performance of each method by estimating its overall accuracy (Ac). Here accuracy is the rate at which each method properly predicts a game's outcome. It is a weighted average of the method's sensitivity (Se) and specificity (Sp), where Se is the proportion of times it predicts Team 1 as the winner given Team 1 actually won and Sp is the proportion of times it predicts Team 1 to lose given Team 1 actually lost. These values are weighted using the prevalence (p), the proportion of times that Team 1 wins the game in reality. The formula for accuracy is

$$\mathrm{Ac} = \mathrm{Se}(p) + \mathrm{Sp}(1 - p). \tag{1}$$

While this and related measures are commonly employed to assess medical diagnostic tests, to date we are aware of no interval estimates for Ac. In order to compare the different methods for predicting the winner of Twenty20 cricket matches, we derived large-sample classical confidence intervals and Bayesian credible sets for Ac. We then calculated the resulting interval estimates for the DL, BGS and DMPO methods using the sample described above. To do so, we collected the final score of Team 1 along with Team 2's score and the number of wickets lost after 5 overs were played, after 10 overs were played, and after 15 overs were played. The target score was calculated for each team in the sample. This target score then determined the winner under the DL, BGS and DMPO methods. This allowed calculation of estimates for Se, Sp and p.

Formulation of the large-sample classical confidence interval for accuracy requires derivation of the variance using the delta method. The delta method addresses a random sample with $E(X^i) = \mu$ and a covariance matrix $E(X^i - \mu)(X^i - \mu)^T = \Sigma$. For a given function g with continuous first partial derivatives and specific value of μ for which $\tau^2 = \nabla^T g(\mu)\Sigma\nabla g(\mu) > 0$, we have $\sqrt{n}(g(\bar{X}) - g(\mu)) \to N(0, \tau^2)$ in distribution. In other words, the delta method allows us to say that accuracy is considered to be normally distributed for a large sample, with the variance derived from

$$\mathrm{var}(\mathrm{Ac}) = \begin{bmatrix} \partial\mathrm{Ac}/\partial\mathrm{Se} \\ \partial\mathrm{Ac}/\partial\mathrm{Sp} \end{bmatrix}^T \begin{bmatrix} \mathrm{var}(\mathrm{Se}) & 0 \\ 0 & \mathrm{var}(\mathrm{Sp}) \end{bmatrix} \begin{bmatrix} \partial\mathrm{Ac}/\partial\mathrm{Se} \\ \partial\mathrm{Ac}/\partial\mathrm{Sp} \end{bmatrix}. \tag{2}$$

The resulting variance is as follows, where n_1 is the number of games Team 1 won and n_2 is the number of games Team 1 lost:

$$\mathrm{var}(\mathrm{Ac}) = \frac{p^2 n_2 \mathrm{Se}(1 - \mathrm{Se}) + (1 - p)^2 n_1 \mathrm{Sp}(1 - \mathrm{Sp})}{n_1 n_2}. \tag{3}$$

The formula for the classical 95% confidence interval for accuracy is

$$\widehat{\mathrm{Ac}} \pm 1.96\sqrt{\widehat{\mathrm{var}(\mathrm{Ac})}}. \tag{4}$$

We then calculate this interval for the DL, BGS and DMPO methods for situations where the game was interrupted after 5 overs, 10 overs and 15 overs.

4. Bayesian credible sets for accuracy

The Bayesian credible set is another method that is used as a means for comparison along with the classical confidence interval. The Bayesian method takes into account the possibility of prior information. The prior information is combined with the data to yield posterior values. Here we assumed binomial data so that the number of games correctly predicted as wins for Team 1, x_{11}, and the number of games correctly predicted as losses, x_{22}, are distributed as

$$x_{11} \sim \text{binomial}(n_1, \text{Se}) \quad \text{and} \quad x_{22} \sim \text{binomial}(n_2, \text{Sp}).$$

We also assumed conjugate beta priors so that

$$\text{Se} \sim \text{beta}(\alpha_{\text{Se}}, \beta_{\text{Se}}) \quad \text{and} \quad \text{Sp} \sim \text{beta}(\alpha_{\text{Sp}}, \beta_{\text{Sp}}),$$

which yielded posterior distributions for Se and Sp of

$$\text{Se}\,|\,\boldsymbol{d} \sim \text{beta}(x_{11}+\alpha_{\text{Se}}, n_1-x_{11}+\beta_{\text{Se}}), \quad \text{Sp}\,|\,\boldsymbol{d} \sim \text{beta}(x_{22}+\alpha_{\text{Sp}}, n_2-x_{22}+\beta_{\text{Sp}}), \tag{5}$$

where $\boldsymbol{d} = (x_{11}, x_{22}, n_1, n_2)$.

We took p to be known and equal to 0.5, implying that there is no advantage to batting first. Likewise, no *a priori* information was available for Se and Sp so we let

$$\alpha_{\text{Se}} = \alpha_{\text{Sp}} = \beta_{\text{Se}} = \beta_{\text{Sp}} = 0.5.$$

Monte Carlo sampling from the posteriors in (5) provided 5000 posterior estimates of Se and of Sp. Plugging these values into (1) estimated the posterior distribution for Ac. The distribution was then used to determine the 95% credible set by determining the 2.5th and the 97.5th percentiles. Table 4 provides the calculated accuracies and the interval for each method evaluated. From the table, we see that DL method is slightly more accurate than BGS and that both DL and BGS methods are superior to DMPO.

	Confidence Interval		Credible Set	
	Estimated Accuracy	Interval	Posterior Accuracy	Interval
DL 5 overs	0.767	(0.692, 0.842)	0.764	(0.686, 0.834)
BGS 5 overs	0.741	(0.665, 0.818)	0.739	(0.659, 0.809)
DMPO 5 overs	0.540	(0.502, 0.579)	0.549	(0.514, 0.590)
DL 10 overs	0.850	(0.788, 0.913)	0.846	(0.777, 0.903)
BGS 10 overs	0.842	(0.779, 0.905)	0.839	(0.770, 0.894)
DMPO 10 overs	0.619	(0.574, 0.664)	0.621	(0.577, 0.679)
DL 15 overs	0.921	(0.872, 0.970)	0.916	(0.857, 0.956)
BGS 15 overs	0.903	(0.850, 0.956)	0.898	(0.835, 0.944)
DMPO 15 overs	0.597	(0.555, 0.640)	0.588	(0.549, 0.643)

Table 4. Classical and Bayesian interval estimates for accuracy.

5. ROC curve analysis of methods

Next we considered receiver operating characteristic (ROC) curves, employed by [Manage et al. 2010] for ODI cricket matches, as another method of comparison. ROC curves present a graphical plot of Se vs. $(1-\text{Sp})$ to determine the best method. A greater area under the curve (AUC) for one method implies that it has higher values of Se together with higher values of Sp and that it is better than other methods with lower areas. To obtain the Se and Sp values for the plot, we needed to classify the thresholds. Here we used the ranking system presented in [Manage et al. 2010]. The rankings are as follows:

IF (Revised target − Actual score of Team 2) < -10
THEN Rank $= 1$ (strongly negative),
IF $-10 \leq$ (Revised target − Actual score of Team 2) < -2
THEN Rank $= 2$ (negative),
IF $-2 \leq$ (Revised target − Actual score of Team 2) ≤ 2
THEN Rank $= 3$ (not clear),
IF $2 <$ (Revised target − Actual score of Team 2) ≤ 10
THEN Rank $= 4$ (positive),
IF (Revised target − Actual score of Team 2) > 10
THEN Rank $= 5$ (strongly positive).

Here, "positive" means a victory for Team 1; in other words, the more positive, the more likely that Team 1 will win the game.

This ranking system was used for the three methods evaluated. The data, after ranking, is presented in Tables 5, 6 and 7.

	Ranking (5 overs)						Ranking (10 overs)						Ranking (15 overs)					
	1	2	3	4	5	Tot.	1	2	3	4	5	Tot.	1	2	3	4	5	Tot.
T1 Wins	24	17	7	6	3	57	28	13	4	8	4	57	30	11	2	2	5	50
T2 Wins	5	10	10	9	29	63	0	3	7	16	35	63	0	3	4	11	44	62
Total	29	27	17	15	32	120	28	16	11	26	39	120	30	14	6	13	49	112

Table 5. DL rankings for games interrupted at 5, 10, and 15 overs.

	Ranking (5 overs)						Ranking (10 overs)						Ranking (15 overs)					
	1	2	3	4	5	Tot.	1	2	3	4	5	Tot.	1	2	3	4	5	Tot.
T1 Wins	20	14	4	12	7	57	26	13	4	2	12	57	28	9	5	1	7	50
T2 Wins	1	8	5	12	37	63	0	1	5	12	45	63	0	1	3	10	48	62
Total	21	22	9	24	44	120	26	14	9	14	57	120	28	10	8	11	55	112

Table 6. BGS rankings for games interrupted at 5, 10, and 15 overs.

	Ranking (5 overs)						Ranking (10 overs)						Ranking (15 overs)					
	1	2	3	4	5	Tot.	1	2	3	4	5	Tot.	1	2	3	4	5	Tot.
T1 Wins	2	3	2	9	41	57	5	3	4	3	42	57	4	3	1	15	27	50
T2 Wins	0	1	0	1	61	63	0	0	0	1	62	63	0	0	0	1	61	62
Total	2	4	2	10	102	120	5	3	4	4	104	120	4	3	1	16	88	113

Table 7. DMPO rankings for games interrupted at 5, 10, and 15 overs.

	DL	BGS	DMPO
AUC at 5 overs	0.807	0.811	0.628
AUC at 10 overs	0.894	0.852	0.628
AUC at 15 overs	0.919	0.906	0.709

Table 8. AUC for DL, BGS, DMPO for interruptions at 5, 10 and 15 overs.

ROC curves can now be created for each situation. The situations considered here were hypothetical interruptions after 5 overs, 10 overs and 15 overs. The ROC curves are shown in Figure 1 on the next page. In these graphs, higher curves represent better models. We can see that the Duckworth–Lewis and Bhattacharya–Gill–Swartz methods are higher than the DMPO method. Thus, the DMPO does not perform as well when predicting the winner of Twenty20 cricket matches, as in the case of the ODI matches.

The associated AUC values were calculated using the trapezoidal method and are reported in Table 8. That table also supports the conclusion that the DL method is not significantly different from the BGS method when predicting the winner of Twenty20 cricket matches and both are superior to the DMPO method.

6. Conclusions

This article evaluated three methods of predicting the winner of interrupted Twenty20 cricket matches: the Duckworth–Lewis, discounted most productive overs, and Bhattacharya–Gill–Swartz methods. These methods were compared based on both accuracy and AUC determined from ROC curves. In order to estimate accuracy, formulas for both classical confidence intervals and Bayesian credible sets were derived. A classification threshold provided by [Manage et al. 2010] was used to create ROC curves.

Comparison of the Duckworth–Lewis and Bhattacharya–Gill–Swartz methods using accuracy showed that the posterior accuracies after stoppages at 5 overs, 10 overs and 15 overs in the game were slightly higher for the Duckworth–Lewis method in each situation. The interval estimates overlap, however, implying that the true accuracy rate could be equivalent for these two methods. The ROC curve

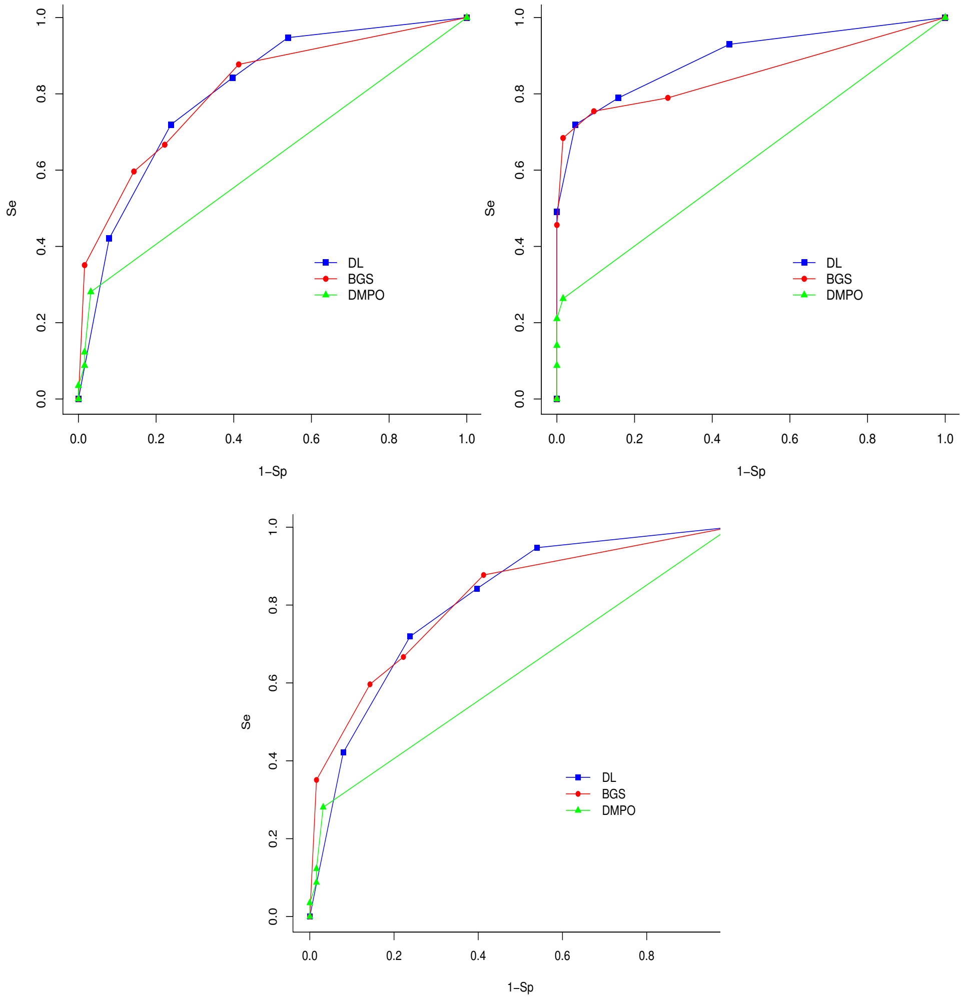


Figure 1. ROC curve for DL, BGS and DMPO for games interrupted at 5, 10, and 15 overs.

method showed similar results, with minimal difference between their AUCs. Thus, we conclude that the Duckworth–Lewis and Bhattacharya–Gill–Swartz methods have comparable success in predicting the winner of interrupted Twenty20 cricket matches. Both outperform the discounted most productive overs method.

Acknowledgements

We thank the anonymous referee for comments that significantly improved this manuscript. We also thank Dr. Ken Smith and the National Science Foundation for their support through grant #DMS-0636528.

References

[Bhattacharya et al. 2010] R. Bhattacharya, P. S. Gill, and T. B. Swartz, "Duckworth–Lewis and Twenty20 cricket", *J. Oper. Res. Soc.* **62**:11 (2010), 1951–1957.

[Duckworth and Lewis 1998] F. Duckworth and A. Lewis, "A fair method for resetting the target in interrupted one-day cricket matches", *J. Oper. Res. Soc.* **49**:3 (1998), 220–227. Zbl 1111.90334

[Manage et al. 2010] A. B. Manage, K. Mallawaarachchi, and K. Wijekularathna, "Receiver operating characteristic (ROC) curves for measuring the quality of decisions in cricket", *J. Quant. Anal. Sports* **6**:2 (2010).

Received: 2012-04-10 Revised: 2013-01-11 Accepted: 2013-03-07

lnm007@shsu.edu
Department of Mathematics and Statistics,
Sam Houston State University, 3307 Legends Mist Drive,
Spring, TX 77386, United States

cnw013@shsu.edu
Department of Mathematics and Statistics,
Sam Houston State University, 11 Mohawk Spur,
Huntsville, TX 77320, United States

sdc017@shsu.edu
Department of Mathematics and Statistics,
Sam Houston State University, 4502 Brazos Bend Drive,
Pearland, TX 77584, United States

jnw009@shsu.edu
Department of Mathematics and Statistics,
Sam Houston State University, PO Box 373,
New Waverly, TX 77358, United States

mholt@shsu.edu
Department of Mathematics and Statistics,
Sam Houston State University, PO Box 2206,
Huntsville, TX 77340, United States

wxb001@shsu.edu
Department of Mathematics and Statistics, Sam Houston State
University, PO Box 2206, Huntsville, TX 77340, United States

INVOLVE 5:3 (2012)
dx.doi.org/10.2140/involve.2012.5.361

Vertex polygons

Candice Nielsen

(Communicated by Colin Adams)

We look at hexagons whose vertex triangles have equal area, and identify necessary conditions for these hexagons to also have vertex quadrilaterals with equal area. We discover a method for creating a hexagon whose vertex quadrilaterals have equal area without necessarily having vertex triangles of equal area. Finally, we generalize the process to build any polygon with an even number of sides to have certain vertex polygons with equal area.

1. Introduction

In the article "Polygons whose vertex triangles have equal area," Harel and Rabin [2003] discuss the properties of polygons with the very special characteristic described in the title. To clarify, the authors offer the following definitions:

Definition 1. A triangle formed using three adjacent vertices of any polygon is called a *vertex triangle*.

Definition 2. A polygon $V_1 V_2 \cdots V_n$ for which all vertex triangles have the same nonzero area is called an *equal-area polygon*.

Harel and Rabin take an algebraic approach, assigning direction and magnitude to each side of the polygon. In this article, we take a geometric approach, using area formulas and triangle congruencies to identify properties of certain polygons.

To extend from triangles, we offer the following definitions:

Definition 3. A polygon of n sides, formed using n adjacent vertices of any m-sided polygon (with $m \geq n$), is called a *vertex n-gon*.

Definition 4. A polygon $V_1 V_2 \cdots V_m$ for which all vertex n-gons have the same nonzero area is called an *equal-n-gon polygon*.

It is clear that every equal-area quadrilateral is also an equal-quadrilateral polygon, since any vertex quadrilateral is the whole quadrilateral. Furthermore, every equal-area pentagon is an equal-quadrilateral polygon because, for P equal to the area of

MSC2010: primary 51N20; secondary 00A99.

Keywords: vertex polygons, vertex triangles, vertex quadrilaterals, geometry, equal area, hexagon, proof by contradiction.

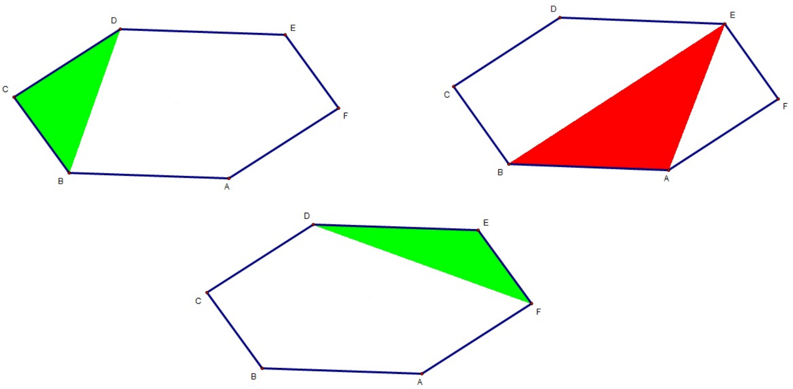


Figure 1. $\triangle BCD$ and $\triangle DEF$ are vertex triangles of hexagon $ABCDEF$, but $\triangle BEA$ is not.

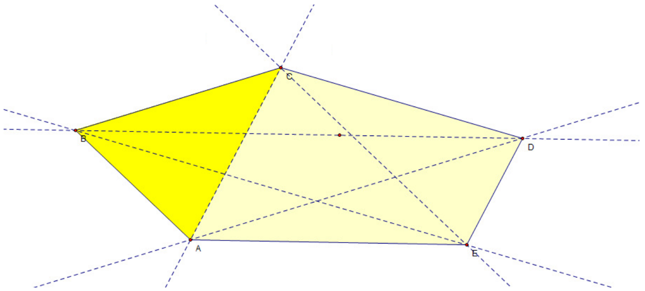


Figure 2. An equal-area pentagon is always an equal-quadrilateral pentagon.

the pentagon, and T equal to the area of any vertex triangle, the area of every vertex quadrilateral is equal to $P - T$ (Figure 2). This means every equal-quadrilateral pentagon is also an equal-area pentagon. For this reason, we begin with hexagons.

2. Equal-area hexagons

The first nontrivial case of the equal-area and equal-quadrilateral polygon is the hexagon. The first task is to construct an equal-area hexagon. We can show that, for any equal-area polygon $V_1 V_2 \cdots V_n$, the line $V_i V_{i+1}$ is parallel to the line $V_{i-1} V_{i+2}$. In other words, each side is parallel to the line formed by the surrounding two vertices.

Proof. Let $V_1 V_2 \cdots V_n$ be an equal-area polygon. Then $\text{Area}(\triangle V_{i-1} V_i V_{i+1}) = \text{Area}(\triangle V_i V_{i+1} V_{i+2})$. Let $b = V_i V_{i+1}$, $h_1 = d(V_{i-1}, V_i V_{i+1})$, $h_2 = d(V_{i+2}, V_i V_{i+1})$. So $\text{Area}(\triangle V_{i-1} V_i V_{i+1}) = \frac{1}{2} b h_1 = \frac{1}{2} b h_2 = \text{Area}(\triangle V_i V_{i+1} V_{i+2})$. Therefore, $h_1 = h_2$ and $V_{i-1} V_{i+2}$ is parallel to $V_i V_{i+1}$. □

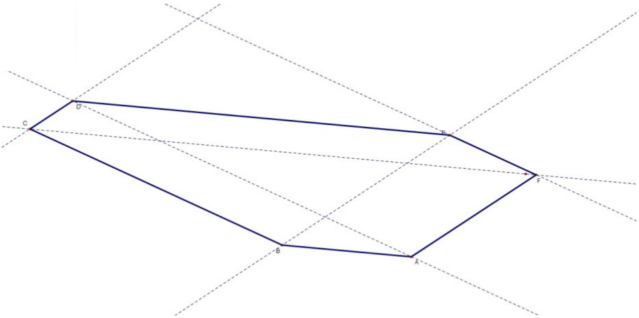


Figure 3. Every equal-area hexagon enjoys parallelism betweenopposite sides and corresponding main diagonals.

With this property, an equal-area hexagon can be uniquely determined by any trapezoid. As we build an equal-area hexagon, it is important to note that the diagonals of the hexagon need not intersect at a single point. This is a key observation as we transform the equal-area hexagon into an equal-quadrilateral hexagon.

Using the hexagon $CDEFGH$ in Figure 3, certain geometric properties arise. First, the sides of the hexagon, along with the diagonals, divide the hexagon into four triangles and three trapezoids. Let us define these as follows:

Definition 5. Let $ABCDEF$ be any hexagon, with $\overline{AD} \cap \overline{BE} = J$, $\overline{BE} \cap \overline{CF} = L$, $\overline{CF} \cap \overline{AD} = K$. The triangle $\triangle JKL$ is called the *center triangle*. The triangles $\triangle ABJ$, $\triangle CKD$, and $\triangle ELF$ are called *interior triangles*, and $BCKJ$, $DELK$, and $FAJL$ are called *interior trapezoids* (see Figure 4).

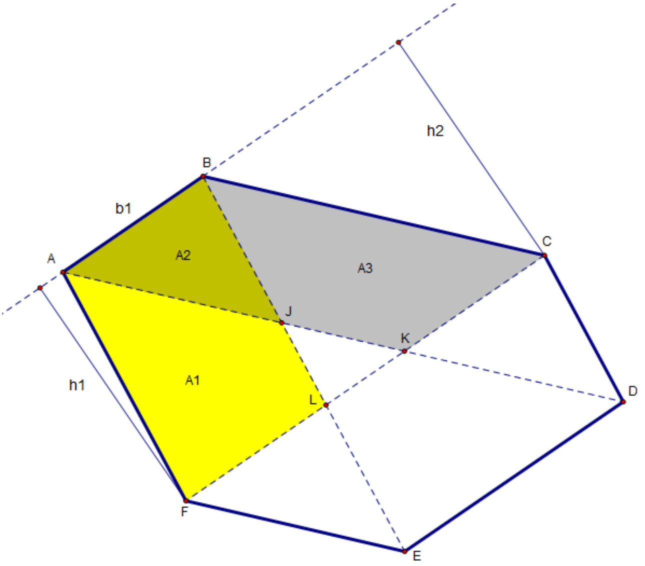


Figure 4. Equal-area hexagon $ABCDEF$ and two of the associated trapezoids.

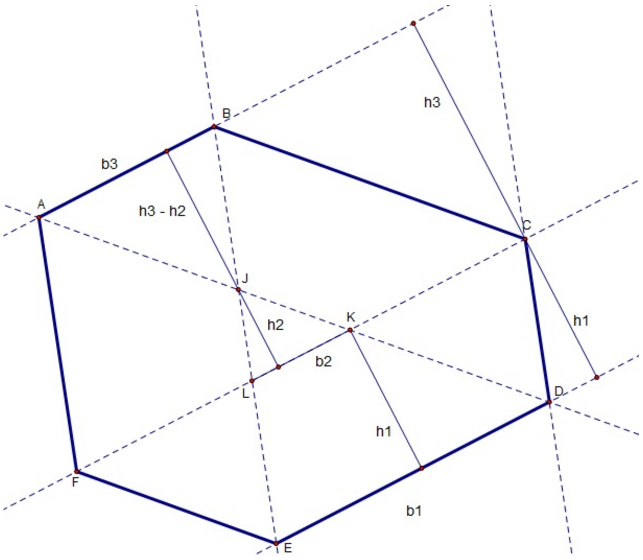


Figure 5. Equal-area, equal-quadrilateral hexagon $ABCDEF$.

Lemma. *For an equal-area hexagon, each interior trapezoid has the same nonzero area and each interior triangle has the same nonzero area.*

Proof. Let $ABCDEF$ be an equal-area hexagon, with $\overline{AD}\cap\overline{BE}=J$, $\overline{BE}\cap\overline{CF}=L$, $\overline{CF}\cap\overline{AD}=K$ (see Figure 5). From the previous proof, $\overline{AF}\parallel\overline{BE}$ and $\overline{AB}\parallel\overline{FL}$, so $ABLF$ is a parallelogram. Likewise, $\overline{AK}\parallel\overline{BC}$ and $\overline{AB}\parallel\overline{CK}$, so $ABCK$ is a parallelogram, and the parallelograms share a base, $\overline{AB}$. Let $b_1=AB$, $h_1=\text{height}(ABLF)=\text{height}(\triangle BAF)$, and $h_2=\text{height}(ABCK)=\text{height}(\triangle ABC)$. Then $\text{Area}(ABLF)=b_1h_1=2\text{Area}(\triangle BAF)$ and $\text{Area}(ABCK)=b_1h_2=2\text{Area}(\triangle ABC)$. Since $ABCDEF$ is an equal-area hexagon, we have $\text{Area}(\triangle BAF)=\text{Area}(\triangle ABC)$, so $\text{Area}(ABLF)=\text{Area}(ABCK)$. Let $A_1=\text{Area}(AJLF)$, $A_2=\text{Area}(\triangle ABJ)$, and $A_3=\text{Area}(BCKJ)$. Then $\text{Area}(ABLF)=A_1+A_2$ and $\text{Area}(ABCK)=A_2+A_3$. This implies that $A_1=A_3$. Similar argument supports that all interior trapezoids have the same nonzero area, as do all interior triangles. □

Definition 6. For any integer $n>1$ and any polygon having n sides with vertices $V_1, V_2, \dots, V_{2n}$, a *true diagonal* has endpoints V_i and V_{i+n}, where $i\in\{1,2,\dots,n\}$.

Theorem 1. *An equal-area hexagon is equal-quadrilateral if and only if all its true diagonals intersect at a single point.*

Proof. Let $ABCDEF$ be an equal-area, equal-quadrilateral hexagon, with the following properties: $\overline{AD}\cap\overline{BE}=J$, $\overline{BE}\cap\overline{CF}=L$, $\overline{CF}\cap\overline{AD}=K$. Suppose, for the sake of

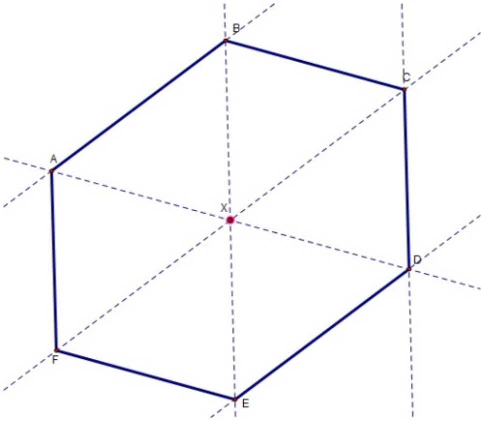


Figure 6. Equal-area hexagon *ABCDEF* with main diagonals intersecting.

contradiction, that J, K, and L are three distinct points. Let A_1 be the area of the interior triangles, A_2 be the area of the interior trapezoids, and A_c be the area of the center triangle. Consider the vertex quadrilaterals *ABCD* and *BCDE*. Since *ABCDEF* is an equal-quadrilateral hexagon, the areas of the vertex quadrilaterals are equal to each other. Thus, $\text{Area}(ABCD) = 2A_1 + A_2 = A_1 + 2A_2 + A_c = \text{Area}(BCDE)$.

Let $b_1 = DE$, and let h_1 equal the height of trapezoid *ELKD*, which is equal to the height of vertex triangle *EDC*.

Let $b_2 = LK$, and let h_2 equal the height of center triangle JKL.

Let $b_3 = AB$ and let h_3 equal the height of vertex triangle ABC, so the height of interior triangle ABJ is $h_3 - h_2$. Then $2A_1 + A_2 = A_1 + 2A_2 + A_c$ implies

$$2\big(\tfrac{1}{2}b_3(h_3 - h_2)\big) + \tfrac{1}{2}(b_1 + b_2)h_1 = \tfrac{1}{2}b_3(h_3 - h_2) + 2\big(\tfrac{1}{2}(b_1 + b_2)h_1\big) + \tfrac{1}{2}b_2h_2$$

This simplifies to

$$b_2h_2 + b_1h_1 + b_2h_1 = b_3h_3 - b_3h_2. \tag{1}$$

Since *ABCDEF* is equal-area, the vertex triangles have the same area, and $\frac{1}{2}b_1h_1 = \frac{1}{2}b_3h_3$, so $b_1h_1 = b_3h_3$ and (1) becomes

$$b_2h_2 + b_2h_1 = -b_3h_2. \tag{2}$$

Since b_2h_2, b_2h_1, b_3h_2 are all positive values, this is a contradiction. Therefore, $J = K = L$, and the diagonals of *ABCDEF* intersect at a single point.

For the other direction, let *ABCDEF* be an equal-area hexagon, satisfying $\overline{AD} \cap \overline{BE} \cap \overline{CF} = X$.

Without loss of generality, consider $\triangle ABX$. Since the height of $\triangle ABX$ is equal to the height of $\triangle ABC$, their areas are equal. Thus, the area of each interior triangle is the area of a vertex triangle. Since all vertex triangles share an equal area, so do the interior triangles. Each vertex quadrilateral is made up of three interior triangles, so each vertex quadrilateral shares an equal area. Therefore, *ABCDEF* is an equal-quadrilateral hexagon. □

3. Equal-quadrilateral hexagons

While constructing an equal-quadrilateral hexagon out of an equal-area hexagon is helpful, the question arose: if a hexagon is equal-quadrilateral, is it necessarily equal-area? We are able to observe, through interior triangle congruencies, that the intersection of the three diagonals is the midpoint of each diagonal. Since the three diagonals are diameters of three concentric circles, we have a new way to construct the equal-quadrilateral hexagon.

Theorem 2. *A hexagon whose true diagonals are diameters of concentric circles is an equal-quadrilateral hexagon.*

Proof. Let AD, BE, CF be diameters of three concentric circles with center X and also be diagonals of hexagon $ABCDEF$ (see Figure 7). Without loss of generality, consider $ABCD$ and $BCDE$. We have $ABCD \cap BCDE = BCDX$. We also have $EX = XB$ and $AX = XD$ because they are radii of the same respective circles. Furthermore, $\angle EXD \cong \angle BXA$ because they are vertical angles. Thus, by the side-angle-side condition, $\triangle EXD \cong \triangle BXA$. Since $BCDX$ is congruent to itself, $ABCD$ and $BCDE$ are congruent, and therefore share an equal, nonzero area. With this argument, every vertex quadrilateral of $ABCDEF$ shares the same, nonzero area. Thus, $ABCDEF$ is an equal-quadrilateral hexagon. □

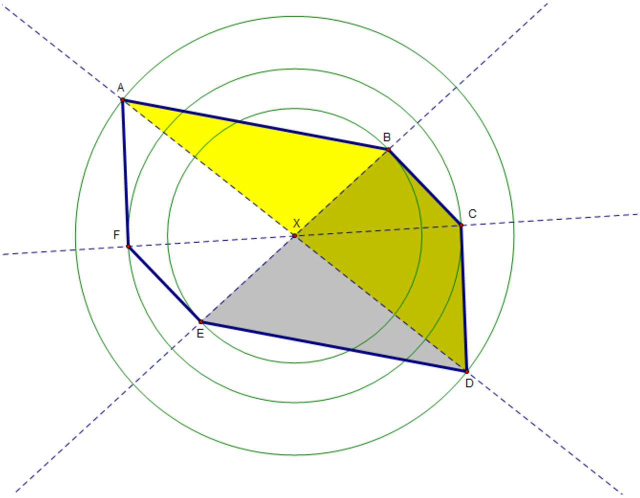


Figure 7. Equal-quadrilateral hexagon *ABCDEF*.

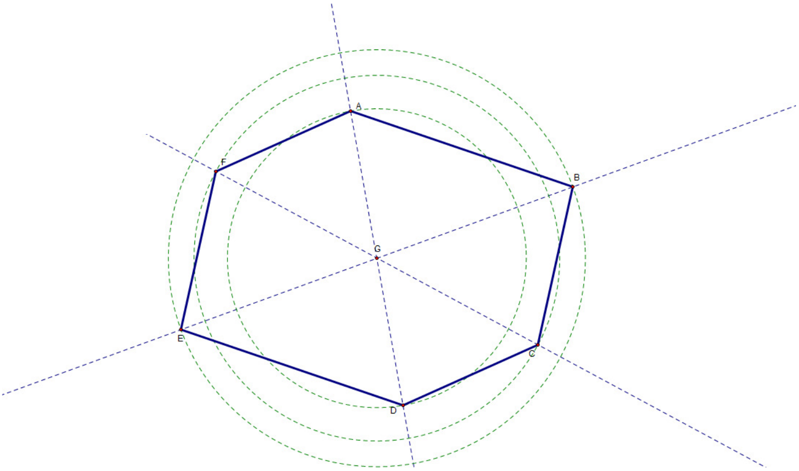


Figure 8. *ABCDEF* is an equal-quadrilateral hexagon, but it is not an equal-area hexagon: Area($\triangle ABC$), Area($\triangle BCD$), and Area($\triangle CDE$) are all different (and each is equal to the area of the symmetrically placed triangle).

To answer the question posed at the beginning of this section, Figure 8 offers a counterexample. All vertex quadrilaterals share an equal area, while the vertex triangles have varying areas.

4. Equal-(n+1)-gon polygons

Corollary. *For any integer $n > 1$, a polygon with $2n$ sides is an equal-$(n+1)$-gon polygon if its true diagonals are diameters of n concentric circles* (see Figure 9).

Proof. Let $n \in Z$, with $n > 1$. Let P_0 be a $2n$-sided polygon constructed using the endpoints of diameters of n concentric circles. Call the center of the circles B, and denote the vertices of P_0 by $V_1, V_2, V_3, \dots, V_{2n}$.

Let P_1 be a polygon with vertices $V_i, V_{i+1}, \dots, V_{i+n}$, and let Area$(P_1) = A_1$. Let P_2 be the polygon with vertices $V_{i+1}, V_{i+2}, \dots, V_{i+n+1}$ and let Area$(P_2) = A_2$. We have $P_1 \cap P_2 =$ polygon$(V_{i+1}, V_{i+2}, \dots, V_{i+n}) \cup \triangle V_{i+1}V_{n+1}B$, which we will call Q_0. Note that Area(Q_0) is equal to itself, so we need only to prove that Area$(P_1 - Q_0) =$ Area$(P_2 - Q_0)$. Since BV_{i+n} and BV_i are radii of the same circle, they are congruent, and likewise for BV_{i+1} and BV_{i+n+1}. Angles V_iBV_{i+1} and $V_{i+n}BV_{i+n+1}$ are congruent because they are vertical angles. Thus, by the side-angle-side formula, the triangles are congruent and therefore have equal area.

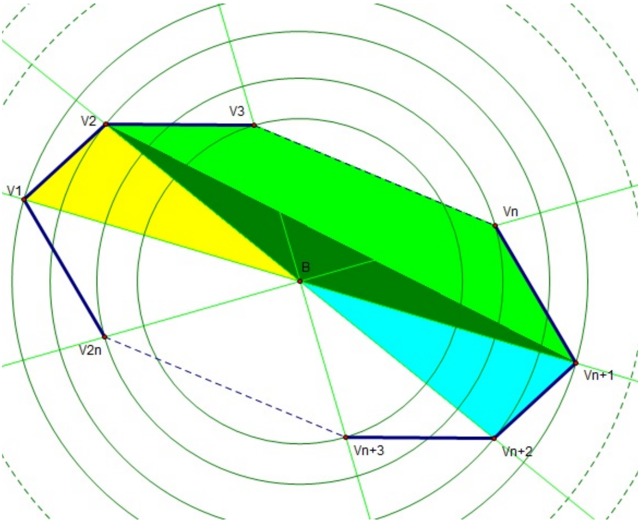


Figure 9. P_1 constructed from diameters of concentric circles.

So Area$(P_1 - Q_0) =$ Area$(P_2 - Q_0)$, and we finally have

$$\begin{aligned}\text{Area}(P_1) &= \text{Area}(Q_0) + \text{Area}(P_1 - Q_0)\\ &= \text{Area}(P_2 - Q_0) + \text{Area}(Q_0) = \text{Area}(P_2).\end{aligned}$$

Therefore, the areas of all vertex $(n+1)$-gons are equal to each other. □

5. Results and open questions

Using known properties of equal-area polygons, we discovered properties of the equal-quadrilateral hexagon. We stated and proved a result that gives necessary conditions for an equal-area hexagon to also be equal-quadrilateral. Finally, we were able to generalize the process of constructing an equal-quadrilateral hexagon to allow construction of any equal-$(n+1)$-gon polygon.

An additional observation on the equal-area hexagon, whether convex or nonconvex, is that the area of the hexagon is equal to the sum of the areas of the vertex triangles. Likewise, the area of any equal-quadrilateral hexagon is twice the area of the vertex quadrilaterals. While this is immediately clear for a convex hexagon, it is not so when the hexagon is nonconvex. Since it is likely the proofs for these observations are simple, they were omitted from this article.

Some questions to consider in extending the idea of equal-n-gon polygons are:

(1) Given an equal-area heptagon, what are the necessary conditions to imply an equal-quadrilateral heptagon? Does equal-quadrilateral imply equal-area in heptagons? If not, how can we construct an equal-quadrilateral heptagon?

(2) Our corollary applies only to polygons with an even number of sides. Given a polygon with an odd number of sides, are there sufficient conditions to ensure vertex polygons of equal area?

Acknowledgements

The author wishes to thank her advisor, Dr. Jon Johnson, for his guidance and support throughout the semester.

References

[Harel and Rabin 2003] G. Harel and J. M. Rabin, "Polygons whose vertex triangles have equal area", *Amer. Math. Monthly* **110**:7 (2003), 606–619. MR 2004e:51021 Zbl 1046.51008

Received: 2012-09-18 Revised: 2012-11-10 Accepted: 2012-12-15

nielsenc@net.elmhurst.edu *Mathematics Department, Elmhurst College,*
Elmhurst, IL 60126, United States

INVOLVE 5:3 (2012)
dx.doi.org/10.2140/involve.2012.5.371

Optimal trees for functions of internal distance

Alex Collins, Fedelis Mutiso and Hua Wang

(Communicated by Jerrold Griggs)

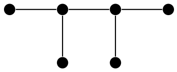
The sum of distances between vertices of a tree has been considered from many aspects. The question of characterizing the extremal trees that maximize or minimize various such "distance-based" graph invariants has been extensively studied. Such invariants include, to name a few, the sum of distances between all pairs of vertices and the sum of distances between all pairs of leaves. With respect to the distances between internal vertices, we provide analogous results that characterize the extremal trees that minimize the value of any nonnegative and nondecreasing function of internal distances among trees with various constraints.

1. Introduction

As a classic example of the distance-based graph invariants, the *Wiener index* [1947] is one of the most well used chemical indices that correlate a chemical compound's structure (the "molecular graph") with experimentally gathered data of the compound's physical-chemical properties such as boiling point, surface pressure, etc. The Wiener index is defined as

$$W(G) = \sum_{\{u,v\}\subseteq V(G)} d(u, v),$$

where $d(u, v)$ is the distance between two vertices u and v and the sum is over all pairs of vertices. For example, the tree shown here has index 29:



The extremal trees that maximize or minimize the Wiener index among general trees [Dobrynin et al. 2001], trees with a given maximum degree [Fischermann et al. 2002], and trees with given degree sequence [Zhang et al. 2008; 2010] have been characterized through various approaches. A general approach was presented dealing with functions of distances between vertices [Schmuck et al. 2012].

MSC2010: primary 05C05, 05C12; secondary 05C30.
Keywords: internal distances, trees, extremal.
This work was partially supported by a grant from the Simons Foundation (#245307 to Hua Wang).

Recently, the *gamma index* [Székely et al. 2011], also known as the *terminal Wiener index* [Gutman et al. 2009], was introduced due to its applications in phylogenetic reconstruction and biochemistry. For a tree T, the gamma index is defined as the sum of distances between all pairs of leaves. It is interesting to note that the star minimizes both the Wiener index and the gamma index among trees of given order. Among trees of a given degree sequence, the "greedy tree" (Definition 1.1) was shown to minimize both the Wiener index [Zhang et al. 2008] and the gamma index [Székely et al. 2011].

Definition 1.1 (greedy trees). With given vertex degrees, the *greedy tree* is achieved through the following *greedy algorithm*:

(i) Label the vertex with the largest degree as v (the root).

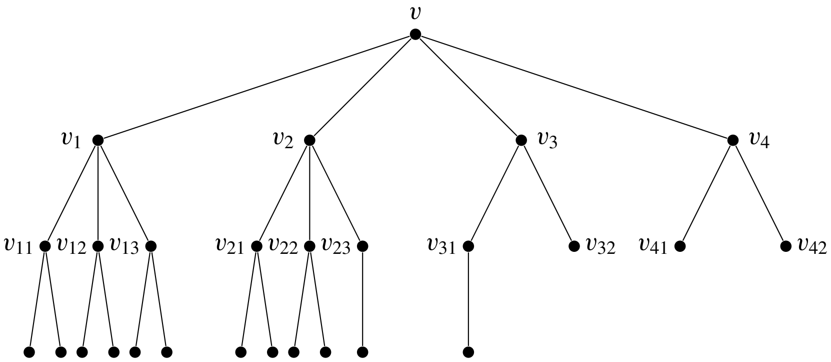
(ii) Label the neighbors of v as $v_1, v_2, \ldots$, and assign the largest degrees available to them such that $\deg(v_1) \geq \deg(v_2) \geq \cdots$.

(iii) Label the neighbors of v_1 (except v) as $v_{11}, v_{12}, \ldots$ such that they take all the largest degrees available and that $\deg(v_{11}) \geq \deg(v_{12}) \geq \cdots$, then do the same for $v_2, v_3, \ldots$.

(iv) Repeat (iii) for all the newly labeled vertices, always starting with the neighbors of the labeled vertex with largest degree whose neighbors are not labeled yet.

For example, here is a greedy tree with degree sequence

$$\{4, 4, 4, 3, 3, 3, 3, 3, 3, 3, 2, 2, 1, \ldots, 1\}.$$



Theorem 1.2 [Schmuck et al. 2012]. *Let $f(x)$ be any nonnegative, nondecreasing function of x. Then the graph invariant*

$$W_f(T) = \sum_{\{u,v\} \subseteq V(T)} f(d(u,v))$$

is minimized by the greedy tree among all trees with given degree sequence.

Theorem 1.2 immediately implies the extremality of the greedy tree regarding many different distance-based graph invariants. Take, for instance, the Wiener index ($f(x) = x$), the hyper-Wiener index ($f(x) = x(x+1)/2$), and the generalized Wiener index ($f(x) = x^{\alpha}$). See [Schmuck et al. 2012] for more details.

Following the Wiener index and the gamma index, a natural next step is to consider the sum of distances between internal vertices. In [Székely and Wang 2005], it was asked if the extremal values of the sums of distances between internal vertices, between leaves, or between internal vertices and leaves can be explored through a similar approach. The sum of distances between internal vertices was brought up again in [Bartlett et al. 2013] and named the *spinal index*:

$$S(T) = \sum_{\{u,v\} \subseteq V(T)-L(T)} d(u, v),$$

where $L(T)$ denotes the set of leaves of T. The extremal trees that maximize or minimize the spinal index have been studied based on the known results regarding the Wiener index [Bartlett et al. 2013]. Similar to $W_f(T)$, it is natural to consider the *spinal function index*, defined as

$$S_f(T) = \sum_{u,v \in V(T)-L(T)} f(d(u, v))$$

for any nonnegative, nondecreasing function f.

The goal of this note is to show that one can provide general statements analogous to Theorem 1.2 and its consequences for $S_f(T)$. By establishing Proposition 2.4, we characterize the trees that minimize the spinal function index among trees with given order and number of leaves (Theorem 3.2), with given degree sequence (Theorem 3.4), as well as with given order and maximum degree (Theorem 3.5).

2. Preliminaries

Our study consists of a combination of techniques in [Bartlett et al. 2013] and [Schmuck et al. 2012]. We first recall the following crucial result, where $p_k(T)$ is the number of pairs (u, v) of vertices such that $d(u, v) \le k$.

Theorem 2.1 [Schmuck et al. 2012]. *Let $d_1 \ge d_2 \ge \cdots \ge d_n$ be positive integers such that $\sum_i d_i = 2(n-1)$, and let k be another arbitrary positive integer. Among all trees with degree sequence $(d_1, d_2, \dots, d_n)$, the greedy tree maximizes $p_k(T)$.*

Remark 2.2. Theorem 2.1 implies Theorem 1.2. Indeed, note that

$$W_f(T) = \sum_{k \ge 0} \bigl(f(k+1) - f(k)\bigr) \bigl| \bigl\{ \{u, v\} \subseteq V(T) \bigm| d(u, v) > k \bigr\} \bigr|,$$

and that $f(k) - f(k-1)$ is nonnegative for all k (we set $f(0) = 0$).

The idea of comparing greedy trees with different degree sequences through "majorization" was used in [Zhang et al. 2012], where the following is defined.

Consider two nonincreasing sequences $\pi = (d_0, \dots, d_{n-1})$, $\pi' = (d'_0, \dots, d'_{n-1})$. If

$$\sum_{i=0}^{k} d_i \le \sum_{i=0}^{k} d'_i \quad \text{for } k = 0, \dots, n-2 \quad \text{and} \quad \sum_{i=0}^{n-1} d_i = \sum_{i=0}^{n-1} d'_i,$$

then π' is said to *majorize* the sequence π, denoted by

$$\pi \lhd \pi'.$$

Lemma 2.3 [Wei 1982]. *Let $\pi = (d_0, \dots, d_{n-1})$ and $\pi' = (d'_0, \dots, d'_{n-1})$ be two nonincreasing graphic degree sequences. If $\pi \lhd \pi'$, then there exists a series of graphic degree sequences $\pi_1, \dots, \pi_m$ such that $\pi \lhd \pi_1 \lhd \cdots \lhd \pi_m \lhd \pi'$, where π_i and π_{i+1} differ at exactly two entries, say d_j (d'_j) and d_l (d'_l) of π_i (π_{i+1}), with $d'_j = d_j + 1$, $d'_l = d_l - 1$ and $j < l$.*

With Lemma 2.3, the following can be shown in a way similar to Theorem 2.4 in [Zhang et al. 2012].

Proposition 2.4. *For two different degree sequences π and π', if $\pi \lhd \pi'$, then*

$$p_k(T^*_\pi) \le p_k(T^*_{\pi'})$$

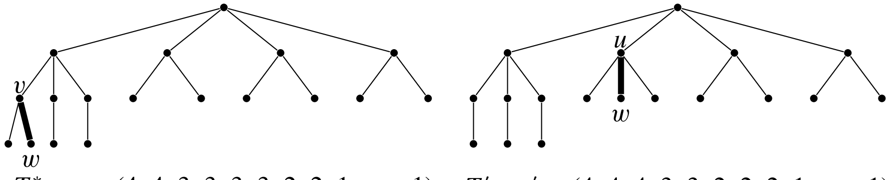
for any $k \ge 1$ where T^*_π and $T^*_{\pi'}$ are the greedy trees with degree sequences π and π', respectively.

Proof. By Lemma 2.3, it is sufficient to show the statement for degree sequences

$$\pi = (d_0, \dots d_{n-1}) \lhd (d'_0, \dots, d'_{n-1}) = \pi'$$

that differ only at the j-th and l-th entries with $d'_j = d_j + 1$, $d'_l = d_l - 1$ for some $j < l$.

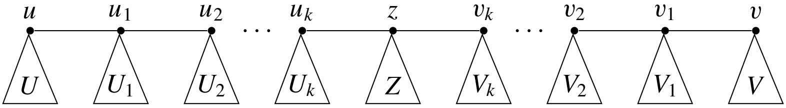
Let T'_π be the tree constructed from T^*_π by removing the edge vw and adding an edge uw, where u and v are the vertices corresponding to d_j and d_l, respectively, and w is a child of v:



T^*_π, $\pi = (4, 4, 3, 3, 3, 3, 2, 2, 1, \dots, 1)$ T'_π, $\pi' = (4, 4, 4, 3, 3, 2, 2, 2, 1, \dots, 1)$

Let T' be the tree obtained from T^*_π after removing w and its descendants. Then the next claim follows from the structure of the greedy tree T^*_π (see, for instance, [Wang 2008; Zhang et al. 2008; 2012]).

Claim 2.5. *Let the path from u to v be $uu_1u_2\cdots u_m(z)v_m\cdots v_2v_1v$, where the existence of z depends on the parity of $d(u, v)$. Let U, U_i, V, V_i, Z denote the component containing u, u_i, v, v_i, z, respectively, after removing the edges on this path from T':*



Then $p_k(U, u) \geq p_k(V, v)$ and $p_k(U_i, u_i) \geq p_k(V_i, v_i)$ for any $1 \leq i \leq m$ and any $k \geq 1$. Here $p_k(T, x)$ denotes the number of vertices $y \in V(T)$ such that $d(x, y) \leq k$.

In particular, Claim 2.5 implies that, for any $k \geq 1$, there are more vertices within distance k from u in T' than those from v.

Now simple calculations (see [Wang 2008], for instance) show that

$$p_k(T^*_{\pi'}) \geq p_k(T'_\pi) \geq p_k(T^*_\pi) \quad \text{for any } k \geq 1. \qquad \square$$

3. Extremal trees with respect to $S_f(T)$

First note that any tree T that is not a star has at least two internal vertices. Hence $S_f(T) \geq 0$ for any T. The following observation is trivial.

Proposition 3.1. *Among all trees with the same order, the star has the minimal $S_f(T) = 0$.*

As shown in Remark 2.2, we only need to focus on $p_k(T)$ for other more involved cases. In what follows we show that several statements from [Bartlett et al. 2013] can be easily generalized for $S_f(T)$. It is worth pointing out that, with the understanding of the preliminaries (particularly with Proposition 2.4), these results can be obtained in a very similar fashion as [Bartlett et al. 2013].

Theorem 3.2. *For a tree T with given order and number of leaves, let T^*_π denote a greedy tree with degree sequence*

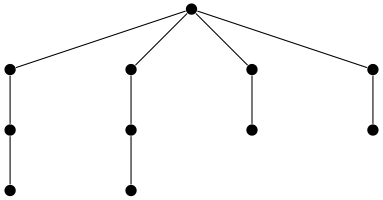
$$\pi = \bigl(|L(T)|, 2, \ldots, 2, \underbrace{1, \ldots, 1}_{|L(T)|\text{ 1's}}\bigr).$$

Then $p_k(T^*_\pi) \geq p_k(T)$ for any $k \geq 1$. Hence $S_f(T)$ is minimized by the same tree by Remark 2.2.

Remark 3.3. Such a tree is called "star-like", achieved by attaching exactly one pendant edge to each of the leaves of a greedy tree with degree sequence

$$(|L(T)|, 2, \ldots, 2, 1, \ldots, 1).$$

Here is an example:



Proof. Consider the subtree T' induced by the internal vertices of T. We have that $p_k(T')$ is maximized by a greedy tree with $|V(T)|-|L(T)|$ vertices and at most $|L(T)|$ leaves (each of the leaves in T' has at least one vertex in $L(T)$ as a neighbor in T).

Among the degree sequences of such trees,

$$\big(|L(T)|, \quad \underbrace{2,\dots,2,}_{|V(T)|-2|L(T)|-1 \text{ 2's}} \quad \underbrace{1,\dots,1}_{|L(T)| \text{ 1's}}\big)$$

majorizes all others. Then T is a greedy tree with degree sequence

$$\big(|L(T)|, \quad \underbrace{2,\dots,2,}_{|V(T)|-|L(T)|-1 \text{ 2's}} \quad \underbrace{1,\dots,1}_{|L(T)| \text{ 1's}}\big). \qquad \square$$

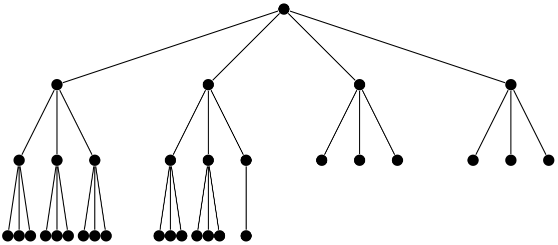
Theorem 3.4. *Among trees with given order and degree sequence, $p_k(T)$ is maximized by the greedy tree. Consequently, $S_f(T)$ is minimized by the greedy tree.*

Proof. First note that with given degree sequence, $|L(T)|$ is determined.

To minimize $S_f(T)$, note that $p_k(T')$ is minimized by a greedy tree with the degree sequence of T'. Let the degree sequence of T be $(d_1, d_2, \dots)$. Then the degree sequence of T' is $(d_1-k_1, d_2-k_2, \dots)$ where $k_i \geq 0$ is the number of leaf-neighbors of the vertex corresponding to the degree d_i. The degree sequence (of T') of this form that majorizes all others is when $k_1 = k_2 = \cdots = k_i = 0$ for i as large as possible. Note that this is the case only when all the vertices (in T) of large degrees have no leaf-neighbors, or in other words, the leaves of T are adjacent only to (as few as possible) internal vertices of the smallest degrees in T. This happens only if T is the greedy tree. Thus the conclusion follows from Proposition 2.4. □

The *complete k-ary tree* with a given maximum degree k is defined in a similar way as the greedy tree, except that the vertices $v, v_1, \dots$ take the maximum degree k until there are not enough vertices (see figure on the next page). As a result, the complete k-ary tree has degree sequence $(k, k, \dots, k, m, 1, \dots, 1)$ for some $1 < m \leq k$.

The extremality of the complete k-ary tree follows in the same way as previous arguments.



A complete 4-ary tree.

Theorem 3.5. *Among trees with given order and maximum degree k, $p_k(T)$ is maximized and $S_f(T)$ is minimized by the complete k-ary tree.*

4. Concluding remarks

We have shown, for any nonnegative, nondecreasing function f, that the sum of $f(d(u, v))$ over all pairs of internal vertices is minimized by the same trees as the ones that minimize the original spinal index. The analogue can be easily obtained for nonincreasing functions. These results, providing a much stronger generalization on this study, were obtained by utilizing tools from previous studies. We also hope that we have illustrated the power of the established approaches in the study of such extremal graphs with respect to distance-based graph invariants.

Acknowledgments

The authors thank the referee for many helpful suggestions that improved the presentation of this note.

References

[Bartlett et al. 2013] M. Bartlett, E. Krop, C. Magnant, F. Mutiso, and H. Wang, "Variations of distance-based invariants of trees", *J. Combin. Math. Combin. Comput.* (2013). To appear.

[Dobrynin et al. 2001] A. A. Dobrynin, R. Entringer, and I. Gutman, "Wiener index of trees: theory and applications", *Acta Appl. Math.* **66**:3 (2001), 211–249. MR 2002i:05035 Zbl 0982.05044

[Fischermann et al. 2002] M. Fischermann, A. Hoffmann, D. Rautenbach, L. Székely, and L. Volkmann, "Wiener index versus maximum degree in trees", *Discrete Appl. Math.* **122**:1-3 (2002), 127–137. MR 2003d:05061 Zbl 0993.05061

[Gutman et al. 2009] I. Gutman, B. Furtula, and M. Petrović, "Terminal Wiener index", *J. Math. Chem.* **46**:2 (2009), 522–531. MR 2011e:05075 Zbl 05601386

[Schmuck et al. 2012] N. S. Schmuck, S. G. Wagner, and H. Wang, "Greedy trees, caterpillars, and Wiener-type graph invariants", *MATCH Commun. Math. Comput. Chem.* **68**:1 (2012), 273–292. MR 2986487

[Székely and Wang 2005] L. A. Székely and H. Wang, "On subtrees of trees", *Adv. in Appl. Math.* **34**:1 (2005), 138–155. MR 2005h:05050 Zbl 1153.05019

[Székely et al. 2011] L. A. Székely, H. Wang, and T. Wu, "The sum of the distances between the leaves of a tree and the 'semi-regular' property", *Discrete Math.* **311**:13 (2011), 1197–1203. MR 2012d:05106 Zbl 1222.05027

[Wang 2008] H. Wang, "The extremal values of the Wiener index of a tree with given degree sequence", *Discrete Appl. Math.* **156**:14 (2008), 2647–2654. MR 2009i:05076 Zbl 1155.05020

[Wei 1982] W. D. Wei, "The class $\mathfrak{A}(R, S)$ of (0, 1)-matrices", *Discrete Math.* **39**:3 (1982), 301–305. MR 84j:05029 Zbl 0484.15015

[Wiener 1947] H. Wiener, "Structural determination of paraffin boiling points", *Journal of the American Chemical Society* **69**:1 (1947), 17–20.

[Zhang et al. 2008] X.-D. Zhang, Q.-Y. Xiang, L.-Q. Xu, and R.-Y. Pan, "The Wiener index of trees with given degree sequences", *MATCH Commun. Math. Comput. Chem.* **60**:2 (2008), 623–644. MR 2009i:05078 Zbl 1195.05022

[Zhang et al. 2010] X.-D. Zhang, Y. Liu, and M.-X. Han, "Maximum Wiener index of trees with given degree sequence", *MATCH Commun. Math. Comput. Chem.* **64**:3 (2010), 661–682. MR 2012d:05130 Zbl 06124037

[Zhang et al. 2012] X.-M. Zhang, X.-D. Zhang, D. Gray, and H. Wang, "The number of subtrees of trees with given degree sequence", *J. Graph Theory* (2012).

Received: 2012-11-05 Revised: 2013-03-10 Accepted: 2013-03-30

acollins38@gsu.edu
Department of Mathematics and Statistics,
Georgia State University, Atlanta, GA 30303, United States

fm00344@georgiasouthern.edu
Department of Mathematical Sciences, Georgia Southern University, Statesboro, GA 30460, United States

hwang@georgiasouthern.edu
Department of Mathematical Sciences, Georgia Southern University, Statesboro, GA 30460, 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 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 BibTeX 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 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

2012 vol. 5 no. 3