involve

a journal of mathematics

msp

involve

msp.org/involve

msp

The influence of education
in reducing the HIV epidemic

Renee Margevicius and Hem Raj Joshi

(Communicated by Suzanne Lenhart)

We use an SIRE (susceptible, infected, removed, and education) model to study
and evaluate the effectiveness of Uganda's education campaigns from the last
25 years in reducing the prevalence of AIDS and HIV infection. We divide
the susceptible class into four subgroups with different infection rates due to
their differing beliefs on sexual conduct. We use data from Uganda about the
epidemic and educational influences to help estimate the infection rates, and then
we simulate the model and compare our results to real data from 1996–2007.

1. Introduction

HIV is a slow working virus that often causes AIDS, in which the immune system
begins to fail. The disease is transmitted by mother to child at birth, or by sharing
needles, or through unsafe sex. At the time of this writing, an estimated 34 million
people were living with HIV throughout the world [WHO 2010], about two-thirds
of them in sub-Saharan Africa.

Uganda, a country located in that region, has had a major influence in HIV
prevention. In 1987, the Ugandan government created a campaign called ABC,
standing for *abstinence*, *being faithful*, and *use of condoms*, to promote ways
of preventing the spread of the virus through safer sexual behavior [Green et al.
2002; 2006]. The first part of the campaign, *abstinence*, promotes no sex until
marriage. The *being faithful* portion supports those couples that only practice sex
with one partner. Lastly, the *use of condoms* promotes safe sex for those with
multiple partners. This three-pronged approach mirrors the recommendations of
international organizations created throughout the world to help educate people on
HIV/AIDS and slow its spread. But Uganda has been more successful than most
countries; throughout the nineties, the prevalence of HIV in Uganda fell, and many
observers credit this to the ABC prevention campaign. Over the last ten years the
incidence of HIV/AIDS in Uganda has largely stabilized [Uganda 2010], even as the

country shifted its prevention policy away from ABC and towards abstinence-only programs, which many experts believe may lead to a rise in risky behavior.

Our goal in this paper is to model the effects of education on the dynamics of the HIV epidemic. In [Joshi et al. 2008], we modified the basic *susceptible, infected, and removed* (SIR) model to include an *education* class (we call the new model an SIRE model). The education class represents the proportion of organizations that are involved in spreading the ABC campaign. We split the susceptible class into three subclasses: the general susceptibles S who do not change their behavior due to the campaign, a class S_{AB} of susceptibles who have been influenced by the *abstinence* and *being faithful* portions, and a class S_C of susceptibles who begin to use *condoms* due to the campaign. Here we extend that work by further dividing S_{AB} into two subclasses S_A and S_B, consisting of those susceptibles who have chosen to practice abstinence and those who have chosen to be faithful to one partner as a result of the campaign.

We collected information such as the history of the HIV epidemic, government statistics, and behavioral records. We used this information in our SIRE model, a system of ordinary differential equations, in order to explain the effects of the ABC campaign. After estimating the parameters using collected data, we simulated the model using MATLAB. The educational influences should cause infection and HIV-related death rates to slow down.

The outline of this paper is as follows. Section 2 will provide an overview of the standard SIR model, as well as the modified SIRE model with the educational influences. Section 3 discusses parameter estimates. In Section 4, we simulate the SIRE model and compare the results to collected data. In Section 5, we present future directions and conclusions.

2. Modified SIRE model

A basic SIR model [Edelstein-Keshet 1988], with susceptible, infected, and removed classes, takes the form

$$S' = -\beta SI + b(S+I) - dS, \quad I' = \beta SI - \gamma I, \quad R' = \gamma I, \quad (1)$$

where d is the natural death rate, b is the birth rate, γ is the death rate due to infection, and β is the infection rate.

We will augment the basic SIR model (1) by introducing educational influence. We will take into account changes in behavior of some susceptibles in the adult population only (ages 15–49), since the promotional campaigns are designed to influence adult behavior. As a result of educational influence, our susceptible class will exhibit different behaviors. Figure 1 is schematic diagram of our model and it shows the connectivity of the different classes.
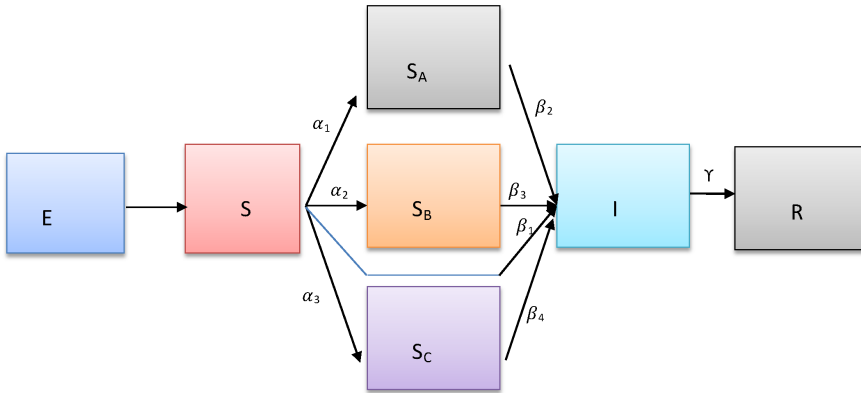
Figure 1. Schematics of the our SIRE model.

The influences of the educational information will break the susceptible popula-
tion into four different types of behavior. The first will be the initial population in
which there is no change in behavior due to education, denoted by S, the general
susceptible population. The next part of the population counts those who have been
influenced to choose abstinence, which will be denoted by S_A. Another type of
change in behavior for the susceptible class will be choosing to be faithful, which
will be indicated by S_B. The last type of behavior change is use of condoms, and
those susceptibles will be denoted by S_C. In addition, an E class is needed to
show the influence of the educational information given about the A, B, and C
type behaviors. This E class causes some members of the susceptible class, S,
to move into the A, B, and C categories. The size of the E class depends on the
fraction of organizations providing the information on HIV. The proportions of
the organizations providing the education on each of the three types of behaviors
causes the split among the A, B, and C categories. Due to influence of E, people
from class S will move into the S_A, S_B, and S_C classes at given rates. In addition,
the entry rate for people into the general susceptible class, and death rates where
people leave, must be taken into consideration.

The infection rate β for each class will vary due to the influence of education
and change in behavior. Thus we will have four different infection rates for the four
different susceptible classes. The infected class I will move to the removed class
R at the rate γ, where the removed class is the number of people who have died
from HIV/AIDS.

For the education class, E, we use a logistic model with a growth rate which
will increase as the number of infective increase. We multiply the growth rate, r,
by the ratio of the populations of the infected to the living.

The following model is an extension of the one in [Joshi et al. 2008]. The main
modification is the use of two separate equations for those with changed behavior

due to abstinence and being faithful. Having two separate susceptible equations for these changes in behavior will result in two different infection rates. These infection rates will be cause members of the respective susceptible subclasses to be added to the infected class and will change the education class. With these parameters and alterations included, our new SIRE model is

$$S' = -\alpha_1 ES - \alpha_2 ES - \alpha_3 ES - \beta_1 SI + b(S + S_A + S_B + S_C + I) - dS,$$

$$S'_A = \alpha_1 ES - \beta_2 S_A I - dS_A \quad (\alpha_1 = 0.02),$$

$$S'_B = \alpha_2 ES - \beta_3 S_B I - dS_B \quad (\alpha_2 = 0.08),$$

$$S'_C = \alpha_3 ES - \beta_4 S_C I - dS_C \quad (\alpha_3 = 0.8),$$

$$I' = \beta_1 SI + \beta_2 S_A I + \beta_3 S_B I + \beta_4 S_C I - \gamma I,$$

$$R' = \gamma I,$$

$$E' = \frac{I}{I + S + S_A + S_B + S_C} rE(1 - E), \tag{2}$$

where α_1, α_2 and α_3 are the transfer rates from S to S_A, S_B, and S_C, respectively. The initial conditions for this system are $S(0)$, $S_A(0)$, $S_B(0)$, $S_C(0)$, $I(0)$, $R(0)$, and $E(0)$. The entering adult rate is b and the general death rate is d. For this case, new adults will enter the general susceptible class S only. Thus, there are four different susceptible classes for which four infection rates are needed: β_1, β_2, β_3, β_4, for S, S_A, S_B, S_C, respectively, as they relate to the infected class I. A proportion of the susceptibles leave the general susceptible class S into S_A, S_B, or S_C when the individuals in class S and the educational campaign class E interact. In addition, when the infected class interacts with the susceptible class, individuals leave according to their rates into the infected class. As a result of HIV, individuals from the infected class leave and are moved into the removed class R with death rate γ.

3. Parameter estimations

The data needed for this model contained information about population, death rates, percentage of adults ages 15–64, the growth of the adult class, adult prevalence rates, and the percent of adult population infected [UNICEF 2010]. In order to determine the organizational estimates for the educational influence rates, we consulted literature, essays, subject matter experts, and surveys. These types of data will influence the relationship between the E and S classes, in addition to the split amongst the A, B, and C behavior types.

The initial conditions for the set of differential equations depend on the data provided. Since the first educational data collected occurred in 1996, we will begin with that year for the model and use the data to determine the initial conditions

for $S(0)$ and $I(0)$. Thus we assume that, prior to 1996, no one followed the A, B, and C type behaviors. In addition, the removed class, R, accumulates the deaths from HIV only. In 1996, the entire population (July 1996 est.) of Uganda was 20,158,176 with adult population comprising 48% [CIA 1997]. Therefore the initial susceptible and infected classes $(S(0) + I(0))$ will have a total of 9,675,924 people for that year. The HIV prevalence rate for adults was estimated to be 12.1%; thus 1,161,110 people are in the infected $(I(0))$ class. As a result, there will be 8,365,991 people in the susceptible $(S(0))$ class. Note that, for $E(0)$, there was an initial estimate of 30% of organizations involved in the ABC campaign. This estimate is an approximation, so the numerical runs will vary. Thus, these are the initial conditions:

$$S(0) = 9.67, \quad S_A(0) = 0, \quad S_B(0) = 0, \quad S_C(0) = 0,$$
$$I(0) = 1.16, \quad R(0) = 0.11, \quad E(0) = 0.30. \tag{3}$$

4. SIRE model simulation

The time span for this model is 12 years (1996–2007). All rates, b, d, r, β_1, β_2, β_3, β_4, γ, were assumed to be constant for all our model simulations. Using data from this time period, we were able to calculate the number of new adults as a percentage of all adults. The entering adult rates for the 12 years were averaged to obtain an entry rate for the susceptible class. The natural death rate was also averaged from UN data, over each five year period. The adults for the general susceptible population had an entering rate $b = 0.055$ and death rate $d = 0.0176$. For γ, we took an average of the death rates due to HIV for a few years and found $\gamma = 0.14$.

As for the parameters, we had to make many assumptions and estimations. For the infection rate parameters, β_1, β_2, β_3, and β_4, we assumed that β_2, β_3, and β_4 were proportional to β_1, so we only needed to determine one infection parameter. We predicted that β_1 was larger than β_2, expecting that the A behavior led to a lower infection rate compared to the general susceptible class. For example, $\beta_2 = 0.01\beta_1$ $(\beta_2 \ll \beta_1)$. As for the infection rate for the B behavior compared to the general susceptible class, we took $\beta_3 = 0.03\beta_1$ $(\beta_2 < \beta_3 \ll \beta_1)$. Lastly, we took the infection rate for the C behavior as $\beta_4 = 0.4\beta_1$ $(\beta_2 < \beta_3 \ll \beta_4 \lll \beta_1)$. We determined the range of values for β_1 that best fit observations, since this infection rate was the hardest to estimate. In addition, we varied r in increments to determine which value, together with β_1, gave a model best fitted for the data.

For the determination of β_1, we first fixed the bounds $0.0001 \leq \beta_1 \leq 0.1$, and let it vary in increments of 0.001, giving 100 values for β_1. Similarly the growth rate r was assumed to lie in the range $0.2 \leq r \leq 2$. We used increments for r of 0.01, giving 180 values. For each pair (β_1, r) we ran the set of differential equations with a MATLAB differential equation solver to give model estimates for the values for

Year	Susceptible	Infected	Removed	Education
1997	9.99	1.046	0.1195	600/1,200
1998	10.42	1.021	0.1205	
1999	10.72	0.975	0.121	
2000	10.96	0.921	0.1215	700/1,200
2001	11.27	0.879	0.120	717/1,200
2002	11.61	0.824	0.118	
2003	12.05	0.807	0.113	
2004	12.47	0.773	0.104	
2005	12.82	0.756	0.089	778/1,200
2006	13.25	0.755	0.084	
2007	14.22	0.754	0.079	

Table 1. Historical data table (population numbers in millions).

each class for each year. These model numbers were then compared to the found data from 1997–2007 [UNAIDS 2009; WHO 2010; Joshi et al. 2008; AVERT 2010; Uganda 2010; UNICEF 2010]. The data points used are shown in Table 1.

We next show our results after running the simulations against our data. Figure 2 represents the true total susceptible data versus the model prediction. The population for this graph is given in millions. The model for the *S* class has data points close, but a few data points are not as close as the others to the graph. For Figure 3, we graphed the model output alongside the infected data. This model shows similarities
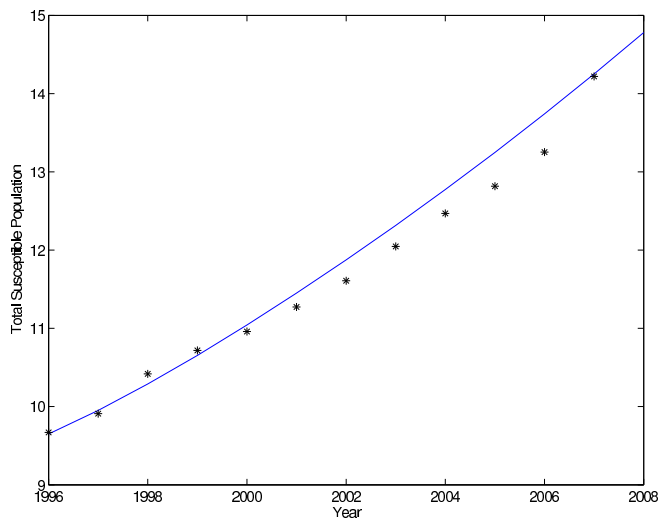
Figure 2. Susceptible population, in millions: model prediction (solid line) and data (∗).
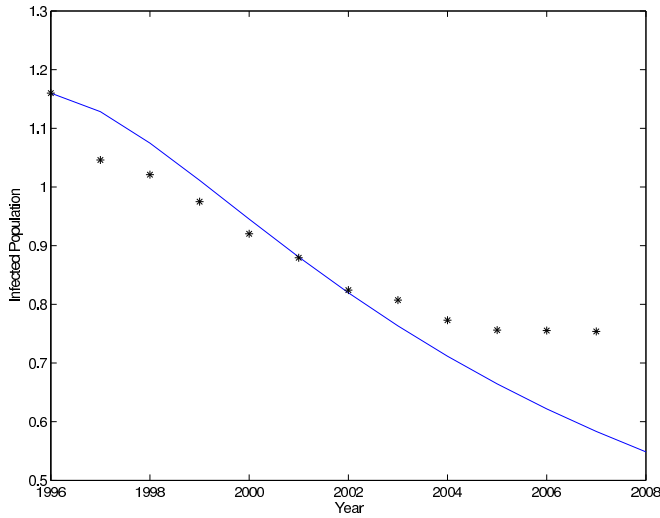
Figure 3. Infected population, in millions: model prediction (solid line) and data (∗).
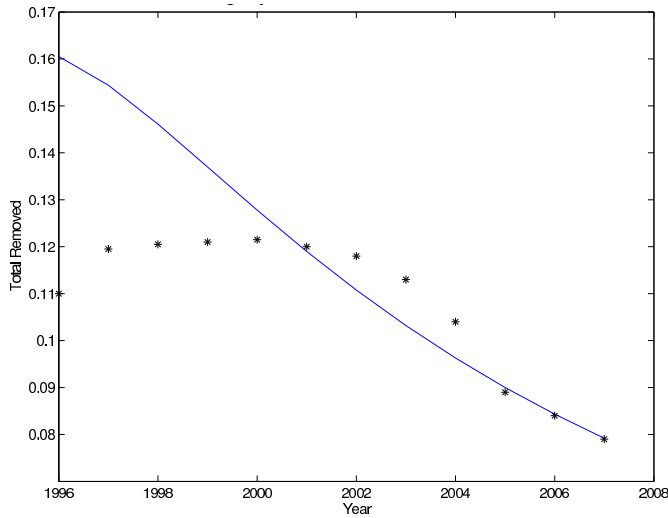
Figure 4. Number of HIV related deaths, in millions: model prediction (solid line) and data (∗).

with the graph. Figure 4 represents the data of deaths per year compared to the model. This model shows a similar shape as the majority of the data points above the model. Figure 5 shows the model for the education class, illustrating fairly close data points to the equation.
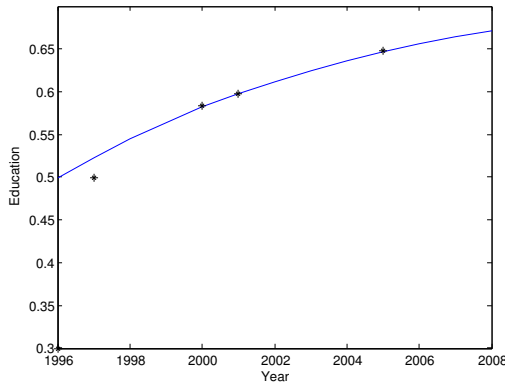
Figure 5. Education influence: model prediction (solid line) and data (∗).

5. Conclusions and future research

This work illustrated altering the susceptible class based on behavior changes due to education. We found that, for the most part, the model's predictions were close to the data. Finding more data and including other features to make this model more realistic is important.

Future research includes refining the S and I classes based on age, gender, and stages of the disease. Past research has shown that females have been more responsive to these educational campaigns than males. Therefore, gender differentiation would be interesting to consider for future modeling efforts. In addition, the involvement of different types of organizations could be studied in variants of the model.

References

[AVERT 2010] AVERT, "HIV and AIDS in Uganda", web page, 2010, http://www.avert.org/aidsuganda.htm.

[CIA 1997] "The world factbook 1996", CIA, 1997, http://www.cia.gov/library/publications/the-world/factbook.

[Edelstein-Keshet 1988] L. Edelstein-Keshet, *Mathematical models in biology*, Random House, New York, 1988. MR 90i:92001 Zbl 0674.92001

[Green et al. 2002] E. Green, V. Nantulya, R. Stoneburner, and J. Stover, "What happened in Uganda?", case study, USAID, 2002, www.unicef.org/lifeskills/files/WhatHappenedInUganda.pdf.

[Green et al. 2006] E. C. Green, D. T. Halperin, V. Nantulya, and J. A. Hogle, "Uganda's HIV prevention success: The role of sexual behavior change and the national response", *AIDS and Behavior* **10**:4 (2006), 335–346.

[Joshi et al. 2008] H. Joshi, S. Lenhart, K. Albright, and K. Gipson, "Modeling the effect of information campaigns on the HIV epidemic in Uganda", *Math. Biosci. Eng.* **5**:4 (2008), 757–770. MR 2478986 Zbl 1154.92037

[Uganda 2010] "UNGASS country progress report: Uganda", progress report, Government of Uganda, 2010, http://tinyurl.com/962kucq.

[UNAIDS 2009] UNAIDS, "Epidemiological fact sheets on HIV/AIDS and sexually transmitted infections. 2009 update", technical report, UNAIDS, 2009.

[UNICEF 2010] "Uganda: Statistics", web page, 2010, http://www.unicef.org/infobycountry/uganda_statistics.html.

[WHO 2010] World Health Organization, "Global summary of the AIDS epidemic", web chart, 2010, http://www.who.int/hiv/data/2011_epi_core_en.png.

margeviciusr1@xavier.edu   *Department of Mathematics and Computer Science, Xavier University, Cincinnati, OH 45207, United States*

joshi@xavier.edu   *Department of Mathematics and Computer Science, Xavier University, Cincinnati, OH 45207, United States*

■msp

On the zeros of $\zeta(s) - c$

Adam Boseman and Sebastian Pauli

(Communicated by Filip Saidak)

Let $\zeta(s)$ be the Riemann zeta function and $z_0 \in \mathbb{C} \setminus \mathbb{R}$ a zero of $\zeta(s)$. We investigate the graphs of the implicit functions $z : [0, 1) \to \mathbb{C}$, with $z(0) = z_0$ given by

$$\zeta(z(c)) - c = 0.$$

We give zero-free regions for $\zeta(s) - c$ where $c \in [0, 1)$.

1. Introduction

For $\sigma = \Re(s) > 1$, the Riemann zeta function can be written as

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

By analytic continuation, $\zeta(s)$ may be extended to the whole complex plane, with the exception of the simple pole $s = 1$. This analytic continuation is characterized by the functional equation

$$\zeta(1 - s) = 2\Gamma(s)\zeta(s)(2\pi)^{-s} \cos \frac{s\pi}{2}. \tag{2}$$

The existence of a class of zeros of the form $-2n$, $n \in \mathbb{N}$, follows directly from the functional equation. These zeros are called trivial. The Riemann hypothesis states that all nontrivial zeros of $\zeta(s)$ are located on the critical line $\sigma = \frac{1}{2}$.

In order to understand the Riemann zeta function better, various mathematicians have investigated the behavior of its derivatives. Speiser [1935] showed that the Riemann hypothesis is equivalent to $\zeta'(s)$ having no zeros for $0 < \Re(s) < \frac{1}{2}$.

Spira [1965] computed zeros of the first and second derivative of $\zeta(s)$ and noticed that they occur in pairs. Skorokhodov [2003] went further in his computation and noticed that the zeros of derivatives seem to form chains; that is, for each zero s_k of $\zeta^{(k)}(s)$ there is a corresponding zero s_{k+1} of $\zeta^{(k+1)}(s)$. For sufficiently large k, the existence of these chains is a direct consequence of the following theorem.

Theorem 1 [Binder, Pauli and Saidak 2013]. *Let $u \in \mathbb{R}^{>0}$ be a solution of*

$$1 - \frac{1}{e^u - 1} - \frac{1}{e^u}\left(1 + \frac{1}{u}\right) \geq 0.$$

Let $M \in \mathbb{N}$, $M \geq 2$, and $j \in \mathbb{Z}$. Let

$$q_M := \log \frac{\log M}{\log(M+1)} \bigg/ \log \frac{M}{M+1}.$$

If there is $k \in \mathbb{N}$ with

$$q_{M+1}k + (M+2)u \leq q_M k - (M+1)u,$$

then each rectangle $R_j \subset S_M^k$, consisting of all $s = \sigma + it$ with

$$q_M k - (M+1)u < \sigma < q_M k + (M+1)u$$

and

$$\frac{2\pi j}{\log(M+1) - \log M} < t < \frac{2\pi(j+1)}{\log(M+1) - \log M},$$

contains exactly one zero of $\zeta^{(k)}(s)$. This zero is simple.

The existence of the chains of zeros of derivatives can be seen as follows. For a given $M \in \mathbb{N}$, $M \geq 2$ there is $K \in \mathbb{N}$ such that $q_{M+1}k + (M+2)u \leq q_M k - (M+1)u$ for all $k \geq K$. By Theorem 1, for each $k \geq K$ and each $j \in \mathbb{Z}$ there is exactly one zero in a rectangular region given by M, k, and j. Again by Theorem 1 there exists a unique corresponding zero of $\zeta^{(k+1)}(s)$ in the rectangular region given by M, $k+1$, and j, which can be obtained by shifting the first region to the right (and stretching it horizontally). This shows the existence of a chain of zeros of $\zeta^{(K)}(s)$, $\zeta^{(K+1)}(s)$, $\zeta^{(K+2)}(s)$, ...

Skorokhodov also noticed that the zeros of $\zeta(s) - 1$ can be regarded as the first points in these chains, and that there are curves from some zeros of $\zeta(s)$ to these points given by the zeros of $\zeta(s) - c$ for $c \in [0, 1)$ (see Figure 1).

The curves of zeros $s(c)$ of $\zeta(s) - c$ for $c \in [0, 1)$ either end at a zero of $\zeta(s) - 1$ or go off to the left approaching their asymptote

$$t = \Re(s) = \frac{(2m+1)\pi}{\log 2},$$

for some $m \in \mathbb{Z}$ as $\sigma = \Re(s)$ approaches infinity. If each zero of $\zeta(s) - 1$ indeed corresponded to a zero of $\zeta'(s)$, $\zeta''(s)$, $\zeta'''(s)$, ..., then some zeros of $\zeta(s)$ would not correspond to zeros with derivatives, namely those from which the paths of zeros of $\zeta(s) - c$ for $c \in [0, 1)$ go off to the right.

This agrees with the formulas for the number of nontrivial zeros of $\zeta(s)$ and $\zeta^{(k)}(s)$. Namely, let $N(T)$ and $N_k(T)$ denote the number of such zeros ρ with
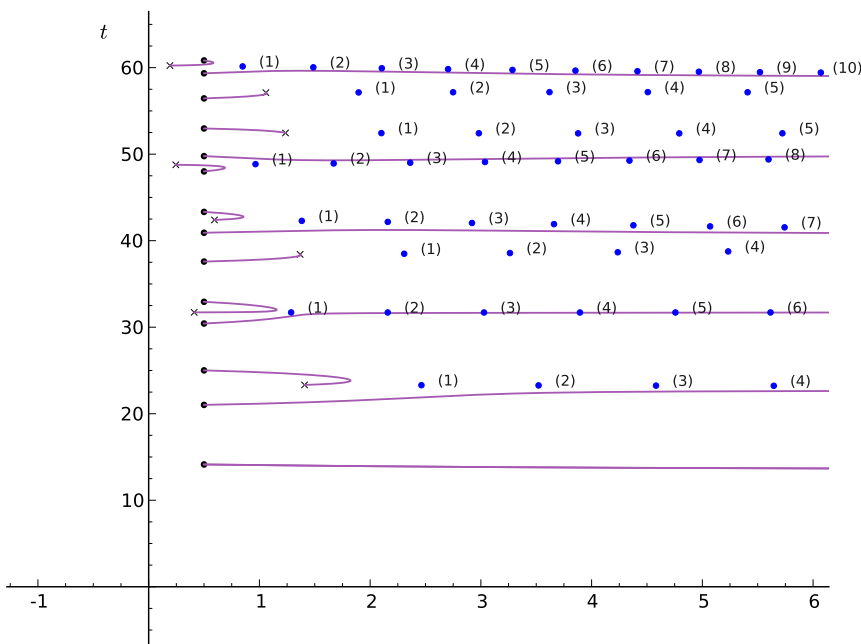
Figure 1. Zeros of derivatives of $\zeta^{(k)}(s)$ (denoted by $\bullet^{(k)}$) and the paths from zeros of $\zeta(s)$ (denoted by $\bullet$) to the zeros of $\zeta(s) - 1$ (denoted by $\times$).

$0 \leq \Im(\rho) \leq T$ of $\zeta(s)$ and $\zeta^{(k)}(s)$, respectively. The classical Riemann–von Mangoldt formula [Landau 1974] states that

$$N(T) = \frac{T}{2\pi} \log \frac{T}{2\pi} - \frac{T}{2\pi} + O(\log T), \tag{3}$$

and according to Berndt [1970], we have

$$N_k(T) = N(T) - \frac{T \log 2}{2\pi} + O(\log T). \tag{4}$$

So there are about $(T \log 2)/2\pi$ fewer zeros of $\zeta^{(k)}(s)$ with imaginary part less than T than there are of $\zeta(s)$, which is also about the number of paths of zeros of $\zeta(s) - c$ with imaginary part less than T that go off to the right.

The aim of this paper is to describe better the behavior of paths of zeros of $\zeta(s) - c = 0$ for $c \in [0, 1)$ by finding new zero-free regions for the functions $\zeta(s) - c$. Our results are summarized in Figure 2. Clearly, the zeros of $\zeta(s) - c$ lie on the real lines of $\zeta(s)$, that is, the lines on which $\Im(\zeta(s)) = 0$. A review of some results about these lines in Section 2 is followed by the derivation of the zero-free
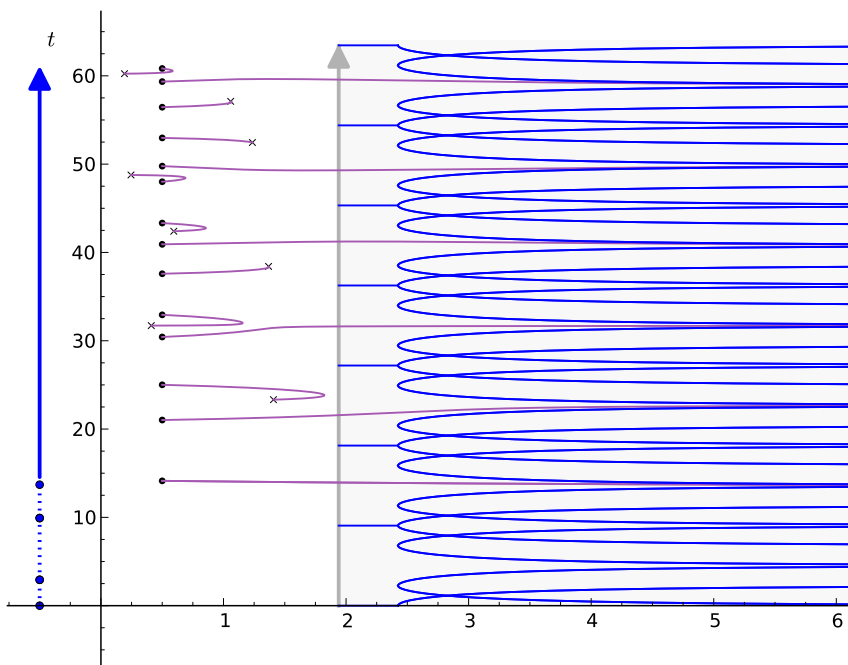
Figure 2. The paths from zeros of $\zeta(s)$ (denoted by •) to the zeros of $\zeta(s) - 1$ (denoted by ×), the barrier on the left (denoted by ↑), the zeros of $\Im\left(\zeta\left(-\frac{1}{2} + it\right)\right)$ with $0 \leq t < 13.7$ (denoted by •), the borders of zero-free regions of $\zeta(s) - c$ for $c \in [0, 1)$ (denoted by blue lines), and the zero-free region of $\zeta(s) - 1$ on the right in gray.

regions for $\zeta(s) - c$ on the right half-plane (Section 3) and the vertical boundary for the zeros of $\zeta(s) - 1$ for $\Re(s) = \frac{1}{2}$ (Section 4).

2. Real lines

Obviously the solutions of the equations $\zeta(s) - c = 0$ where $c \in [0, 1)$ are on the level lines with $\Im(\zeta(s)) = 0$, called real lines. Most of the results described here go back to the work of Speiser and his student Utzinger [Speiser 1935]. Plots of the behavior of the real (and imaginary) lines and some further discussion can be found in [Arias-de-Reyna 2005].

Because the term $1 + 2^{-s}$ dominates the infinite series $\zeta(s) = \sum_{i=0}^{\infty}(1/n^s)$ for $\sigma = \Re(s) > 3$, the real lines have asymptotes $t = j\pi/\log 2$ for $j \in \mathbb{Z}$. On the real lines with asymptote $t = 2m\pi/\log 2$ ($m \in \mathbb{Z}$) the function $\zeta(s)$ approaches 1 from above, while on the real lines with asymptote $t = (2m + 1)\pi/\log 2$ ($m \in \mathbb{Z}$) the function $\zeta(s)$ approaches 1 from below. The zero-free regions for $\zeta(s) - c = 0$

where $c \in [0, 1)$ narrow around these asymptotes as σ increases — see Lemma 4 and Lemma 3.

As $\zeta(s)$ is a meromorphic function, no two of these real lines can cross where $\zeta'(s) \neq 0$. Zero-free regions for $\zeta'(s)$ have been found on the left of the critical line for $\Im(s) \neq 0$ and $\Re(s) < 0$ [Levinson and Montgomery 1974, Theorem 9] ($\Re(s) < \frac{1}{2}$ under the Riemann hypothesis [Speiser 1935]) and on the right of the critical line for $\sigma > 2.94$ [Skorokhodov 2003, Theorem 2]. Indeed, the only point where two real lines coming from the right cross is the first real zero of $\zeta'(s)$ at $s \approx -2.7172628292$ [Speiser 1935]. Here the lines with asymptotes $t = 2\pi/\log 2$ and $t = -2\pi/\log 2$ intersect the real axis.

The lines coming from the right continue to the left at least until $\sigma = 1.95$ (compare Lemma 5). If one of the lines coming from the right did not cross the strip $-1 \leq \sigma \leq 2$, it would have go up towards infinity. Because no two real lines coming from the right intersect, all following lines would have to do the same. This would contradict the estimate

$$\Im\left(\int_{2+Ti}^{-1+Ti} \frac{\zeta'(s)}{\zeta(s)} \, ds\right) = O(\log T)$$

used in the proof of the Riemann–von Mangoldt formula (3). Thus all real lines coming from the right cross the strip $-1 \leq \sigma \leq 2$ [Speiser 1935].

Hence the zeros of $\zeta(s) - c = 0$, where $c \in [0, 1)$, are either on the real lines described above or on real lines that enter the critical strip from the left half-plane and then curve back to the left half-plane. The lines coming from the left half-plane are the lines on which $\zeta(s) - 1$ is 0. By Proposition 7, we have $\left|\zeta\left(-\frac{1}{2} + it\right)\right| > 1$ for $t \geq 13.7$. Furthermore, for $0 < t < 13.7$, there are only two points where $\Re\left(\zeta\left(-\frac{1}{2} + it\right)\right) = 0$, that is, where the real lines with asymptote $t = 2\pi/\log 2$ and $t = 3\pi/\log 2$ cross the line $\sigma = -\frac{1}{2}$ (see Remark 8). It follows that each of these lines coming from the left contains a zero of $\zeta(s)$ and a zero of $\zeta(s) - 1$ on the left of $\sigma = -\frac{1}{2}$. It is well-known that the real part of the zeros of $\zeta(s)$ is between 0 and 1, and equals $\frac{1}{2}$ if one assumes the Riemann hypothesis. An upper bound for the real part zeros of $\zeta(s) - 1$ was given by Skorokhodov [2003]; see Lemma 2 below.

3. Zero-free regions for $\zeta(s) - c$ on the right

A right bound $\sigma = 3$ for the zeros of $\zeta(s) - 1$ can easily be obtained with the triangle inequality and an estimate for $\zeta(\sigma) - 1/2^\sigma - 1$. Skorokhodov was able to get a better bound by applying the triangle inequality to a real-valued function that only considers terms of the zeta function with n odd.

Lemma 2 [Skorokhodov 2003]. *The function $\zeta(s)$ is distinct from unity at $\sigma \in (\sigma_0, \infty)$, where*

$$\sigma_0 = 1.940101683745\ldots$$

is the zero of the function

$$f(\sigma) = 1 + 2^{-\sigma} - (1 - 2^{-\sigma})\zeta(\sigma), \quad \sigma > 1.$$

For $c \in [0, 1)$ we find zero-free regions of $\zeta(s) - c$ that depend on t. We obtain them by considering the real and imaginary parts of $\zeta(s) - c$ separately.

Lemma 3. *If $c \in [0, 1)$ and $|\sin(t \log 2)| \geq 2^\sigma \zeta(\sigma) - 2^\sigma - 1$, then $\zeta(\sigma + it) - c \neq 0$.*

Proof. We consider the imaginary part of $\zeta(s) - c$ and obtain

$$\left|\Im(\zeta(s) - c)\right| \geq \left|\frac{1}{2^\sigma} \sin(t \log 2)\right| - \left|\sum_{n=3}^{\infty} \frac{1}{n^\sigma}\right|$$

$$= \left|\frac{1}{2^\sigma} \sin(t \log 2)\right| - \left|\zeta(\sigma) - 1 - \frac{1}{2^\sigma}\right|, \qquad (5)$$

which is greater than 0 when

$$\left|\sin(t \log 2)\right| \geq 2^\sigma \zeta(\sigma) - 2^\sigma - 1. \qquad \square$$

Lemma 4. *If $c \in [0, 1)$ and $\cos(t \log 2) \geq 2^\sigma \zeta(\sigma) - 2^\sigma - 1$, then $\zeta(\sigma + it) - c \neq 0$.*

Proof. For the real part of $\zeta(s) - c$ we obtain

$$\Re(\zeta(s) - c) = 1 - c + \frac{1}{2^\sigma} \cos(t \log 2) + \cdots$$

$$\geq \frac{1}{2^\sigma} \cos(t \log 2) - \left(\zeta(\sigma) - 1 - \frac{1}{2^\sigma}\right) \quad \text{assuming } c = 1,$$

which is greater than 0 when

$$\cos(t \log 2) \geq 2^\sigma \zeta(\sigma) - 2^\sigma - 1. \qquad \square$$

These regions can be extended a bit if we restrict ourselves to certain values of t.

Lemma 5. *If $c \in [0, 1)$, $m \in \mathbb{Z}$, and t is fixed at $2\pi m/\log 2$, then $\Re(\zeta(s) - c) \neq 0$ for $\sigma \geq 1.95$.*

Proof. $\Re(\zeta(s) - c) = 1 - c + (1/2^\sigma) \cos(t \log 2) + (1/3^\sigma) \cos(t \log 3) + \cdots$ When t is fixed and $t \log 2 = 2\pi m$, we get

$$\Re(\zeta(s) - c) \geq 1 - c + \sum_{\nu=0}^{\infty} \frac{1}{(2^\nu)^\sigma} - \left(\sum_{n=2}^{\infty} \frac{1}{n^\sigma} - \sum_{\nu=0}^{\infty} \frac{1}{(2^\nu)^\sigma}\right)$$

$$= 2 \sum_{\nu=1}^{\infty} \left(\frac{1}{2^\sigma}\right)^\nu - \zeta(\sigma) = \frac{2}{1 - 1/2^\sigma} - \zeta(\sigma),$$

which is greater than 1 for $\sigma \geq 1.95$. $\qquad\square$

4. Zero-free barrier for $\zeta(s) - c$ on the left

On the left, instead of finding a zero-free region, we find a horizontal line where $|\zeta(s)| > 1$. The line $\sigma = -\frac{1}{2}$ fulfills this condition with the exception of one point.

First we find a lower bound for the absolute value of $\zeta(s)$ where $\sigma = \frac{3}{2}$.

Lemma 6. $\left|\zeta\left(\frac{3}{2} + it\right)\right| > 0.46$ *for all* $t \in \mathbb{R}$.

Proof. To get a lower bound for $|\zeta(s)|$, we use the Euler product. Let P be the set of the first million prime numbers, and consider the expression $\prod_{p \in P} |1 - p^{-s}||\zeta(s)|$. We have

$$\prod_{p \in P} |1 - p^{-s}||\zeta(s)| = \left|1 + \sum_{\substack{p \nmid n \\ p \in P}} \frac{1}{n^s}\right| \geq \left|1 - \left|\sum_{\substack{p \nmid n \\ p \in P}} \frac{1}{n^s}\right|\right|$$

$$\geq 1 - \sum_{\substack{p \nmid n \\ p \in P}} \frac{1}{n^\sigma} = 2 - \prod_{p \in P}(1 - p^{-\sigma})\zeta(\sigma).$$

We also have from the triangle inequality that $|1 - p^{-s}| \leq 1 + p^{-\sigma}$, and thus

$$|\zeta(s)| \geq \frac{2 - \prod_{p \in P}(1 - p^{-\sigma})\zeta(\sigma)}{\prod_{p \in P}(1 + p^{-\sigma})} \geq 0.46 \quad \text{for } \sigma = \frac{3}{2}.$$

So we get $|\zeta(s)| \geq \delta > 0$ for $\sigma = \frac{3}{2}$ and $\delta = 0.46$. $\qquad\square$

Now we can use δ and the functional equation to obtain a barrier for the zeros of $\zeta(s) - c$ on the left.

Proposition 7. $|\zeta(-\frac{1}{2} + it)| > 1$ *for* $t \geq 13.7$.

Proof. By the functional equation,

$$\zeta(1 - s) = 2^{1-s}\pi^{-s}\sin\left(\frac{\pi}{2}(1 - s)\right)\Gamma(s)\zeta(s)$$

$$= 2^{1-s}\pi^{-s}\cos\frac{s\pi}{2}\Gamma(s)\zeta(s).$$

Taking the absolute value of both sides gives

$$|\zeta(1 - s)| = 2^{1-\sigma}\pi^{-\sigma}\left|\cos\frac{s\pi}{2}\right||\Gamma(s)||\zeta(s)|.$$

But

$$\left|\cos\frac{s\pi}{2}\right| = \tfrac{1}{2}\left|e^{-\pi(\sigma i - t)/2} + e^{\pi(t - \sigma i)/2}\right|$$

$$= \tfrac{1}{2}\left|e^{-t\pi/2}(\cos\sigma + i\sin\sigma) + e^{t\pi/2}(\cos\sigma - i\sin\sigma)\right|$$

$$= \tfrac{1}{2}\left|\cos\sigma(e^{t\pi/2} + e^{-t\pi/2}) + i\sin\sigma(e^{-t\pi/2} - e^{t\pi/2})\right|$$

$$= \tfrac{1}{2}\left(\cos^2\sigma(e^{\pi t} + e^{-\pi t} + 2) + \sin^2\sigma(e^{\pi t} + e^{-\pi t} - 2)\right)^{\frac{1}{2}}$$

$$= \tfrac{1}{2}\left(e^{\pi t} + e^{-\pi t} + 2(\cos^2\sigma - \sin^2\sigma)\right)^{\frac{1}{2}}.$$

As $\Gamma(z+1) = z\Gamma(z)$ for $z \in \mathbb{C}$ and as

$$\left|\Gamma\left(\tfrac{1}{2} + it\right)\right| = \sqrt{\pi}\,\mathrm{sech}(\pi t) = \sqrt{\frac{2\pi}{e^{\pi t} + e^{-\pi t}}}$$

for $t \in \mathbb{R}$, we get

$$\left|\Gamma\left(\tfrac{3}{2} + it\right)\right| = \left|\left(\tfrac{1}{2} + it\right)\Gamma\left(\tfrac{1}{2} + it\right)\right| = \sqrt{\tfrac{1}{4} + t^2} \cdot \sqrt{\pi} \cdot \sqrt{\frac{2}{e^{\pi t} + e^{-\pi t}}}.$$

For $\sigma = \tfrac{3}{2}$ we obtain

$$\left|\zeta\left(-\tfrac{1}{2} + it\right)\right| \geq 2^{-0.5}\pi^{-1}\frac{1}{\sqrt{2}}\left(1 + \frac{4\cos^2\left(\tfrac{3}{2}\right) - 2}{e^{\pi t} + e^{-\pi t}}\right) \cdot \sqrt{\tfrac{1}{4} + t^2} \cdot \delta,$$

where the right-hand side is obviously increasing in t. With $\delta > 0.46$, this gives $\left|\zeta\left(\tfrac{1}{2} + it\right)\right| > 1$ for $t \geq 13.7$ by Lemma 6. $\square$

Remark 8. The zeros of $\Im\left(\zeta\left(-\tfrac{1}{2} + it\right)\right)$ with $0 \leq t < 13.7$ are $t_0 = 0$, $t_1 \approx 2.93$, and $t_2 \approx 9.92$, where

$$\zeta\left(-\tfrac{1}{2} + it_0\right) \approx -0.21, \quad \zeta\left(-\tfrac{1}{2} + it_1\right) \approx 0.35, \quad \zeta\left(-\tfrac{1}{2} + it_2\right) \approx 2.03.$$

So the only hole in the barrier is $-\tfrac{1}{2} + it_1$. This is where the real line with asymptote $\pi/\log 2$ crosses the line $\sigma = -\tfrac{1}{2}$.

5. Outlook

In our work, we investigated the behavior of the graphs of the continuous functions $s : [0, 1) \to \mathbb{C}$ defined by the equation $\zeta(s(c)) - c = 0$ and an initial point $s(0)$ (a zero of the zeta function). If $s(1)$ exists, such a graph connects a zero of $\zeta(s)$ to a zero of $\zeta(s) - 1$. The latter zeros are the first points on the conjectured chains of zeros of derivatives.

A similar approach could also be used to investigate the conjectured chains of zeros of the derivatives of $\zeta(s)$. For each zero s_0 of

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

one would consider the implicit function $s : [0, \infty) \to \mathbb{C}$ given by

$$\zeta^{(k)}(s(k)) = (-1)^k \sum_{n=1}^{\infty} \frac{\log^k n}{n^{s(k)}} = 0,$$

with $s(0) = s_0$. This function $s(k)$ should yield the correspondence of zeros of $\zeta^{(k)}(s)$ and $\zeta^{(k+1)}(s)$ for $k \in \mathbb{Z}$, $k \geq 0$ for two zeros which would be connected by $\{s(x) \mid k \leq x \leq k + 1\}$.

Together, the two implicit functions could give more detailed insight into the distribution of the zeros of $\zeta(s)$ by relating it to the distribution of higher derivatives (see Theorem 1). Furthermore it will be interesting to see how the conjectured chains of zeros of the derivatives of $\zeta(s)$ fit in with the universality of $\zeta(s)$ found by Voronin [1975].

Acknowledgements

References

[Arias-de-Reyna 2005] J. Arias-de-Reyna, "X-ray of Riemann zeta-function", preprint, 2005. arXiv math/0309433v1.

[Berndt 1970] B. C. Berndt, "The number of zeros for $\zeta^{(k)}(s)$", *J. London Math. Soc.* (2) **2**:4 (1970), 577–580. MR 42 #1776 Zbl 0203.35503

[Binder et al. 2013] T. Binder, S. Pauli, and F. Saidak, "Zeros of high derivatives of the Riemann zeta function", *Rocky Mountain J. of Math.* (2013). To appear.

[Landau 1974] E. Landau, *Handbuch der Lehre von der Verteilung der Primzahlen*, 3rd ed., Chelsea, New York, 1974. MR 16,904d Zbl 0051.28007

[Levinson and Montgomery 1974] N. Levinson and H. L. Montgomery, "Zeros of the derivatives of the Riemann zeta-function", *Acta Math.* **133** (1974), 49–65. MR 54 #5135 Zbl 0287.10025

[Skorokhodov 2003] S. L. Skorokhodov, "Аппроксимации Паде и численный анализ дзета-функции Римана", *Zh. Vychisl. Mat. Mat. Fiz.* **43**:9 (2003), 1330–1352. Translated as "Padé approximation and numerical analysis for the Riemann ζ-function" in *Comput. Math. Math. Phys.* **43**:9 (2003), 1277–1298. MR 2004h:11113 Zbl 1079.41012

[Speiser 1935] A. Speiser, "Geometrisches zur Riemannschen Zetafunktion", *Math. Ann.* **110**:1 (1935), 514–521. MR 1512953 Zbl 0010.16401

[Spira 1965] R. Spira, "Zero-free regions of $\zeta^{(k)}(s)$", *J. London Math. Soc.* **40** (1965), 677–682. MR 31 #5849 Zbl 0147.30503

[Stein et al. 2009] W. Stein et al., "Sage: open-source mathematics software", 2009, Available at http://www.sagemath.org.

[Voronin 1975] S. M. Voronin, "Теорема об 'универсальности' дзета-функции Римана", *Izv. Akad. Nauk SSSR Ser. Mat.* **39**:3 (1975), 475–486. Translated as "Theorem on the 'universality' of the Riemann zeta-function" in *Math. USSR Izv.* **9**:3 (1975), 443–453. MR 57 #12419 Zbl 0315.10037

a_bosema@uncg.edu   *Joint School of Nanoscience and Nanoengineering,*
                                 The University of North Carolina at Greensboro,
                                 Greensboro, North Carolina 27402, United States

s_pauli@uncg.edu   *Department of Mathematics and Statistics,*
                                 The University of North Carolina at Greensboro,
                                 Greensboro, North Carolina 27402, United States

Dynamic impact of a particle

Jeongho Ahn and Jared R. Wolf

(Communicated by John Baxley)

In this work, we consider a moving particle which drops down onto a stationary rigid foundation and bounces off after its contact. The equation of its motion is formulated by a second-order ordinary differential equation. The particle satisfies the Signorini contact conditions which can be interpreted in terms of complementarity conditions. The existence of weak solutions is shown by using a finite time step and the necessary a priori estimates which allow us to pass to the limit. The uniqueness of the solutions can be proved under some additional assumptions. Conservation of energy is also investigated theoretically and numerically. Numerical solutions are computed via both finite- and infinite-dimensional approaches.

1. Introduction

Contact between two bodies happens in our life everyday. Consider, for example, the contact between a floor and an elastic ball such as a basketball or a volleyball, or contact between a brake pad and a disc of a car's wheel. These contact phenomena may seem to be simple from physical or engineering points of view. However, proving the existence of solutions for these contact models requires very sophisticated mathematical analysis and is a mathematical challenge.

Historically, the study of contact mechanics may have originated with [Hertz 1881], where the physicist analyzed a *static* contact problem of two elastic bodies. Mathematical research on contact problems has become more active since Signorini [1933] formulated the general static contact problem of linearly elastic bodies. Most mathematical research on contact mechanics has focused on *static* or *quasistatic* problems and relatively little research on *dynamic* contact problems has been carried out. This has started to change, as mathematical tools and numerical methods for dynamic contact problems have been developed.

Readers interested in contact problems may refer to the remarkable paper [Stewart 2000] for *rigid-body* dynamics with friction and impact which is described by ordinary differential equations (ODEs) and [Kikuchi and Oden 1988] for contact in *elasticity* which deals with elliptic, parabolic, or hyperbolic types of partial differential equations (PDEs).

The study of one-dimensional contact problems is of considerable importance, in its own right and because it provides a foundation for higher-dimensional problems. There are many one-dimensional dynamic contact models involving vibrating strings, elastic rods, and elastic beams modeled in various ways: Euler–Bernoulli beams (linear), Timoshenko beams, and many kinds of nonlinear beams. Nonlinear Gao beams [Gao 1996] are especially noteworthy, as their model allows for buckling, and their contact problems have recently been the subject of many interesting studies; see [Ahn et al. 2012], for example.

There are many open questions in dynamic contact problems. For example, showing the uniqueness of solutions for rigid dynamics models or dynamic contact models between an elastic body and a rigid foundation with Signorini contact conditions is a challenging problem. In addition, proving the existence of solutions for dynamic contact between a purely elastic body and a rigid foundation over more than three dimensions is still an open question. Indeed, the dynamic contact problem has been studied in [Ahn and Stewart 2009], where the viscosity is added to prove the existence of solutions. Mathematically speaking, inserting the viscosity into the equation of motion is a great idea to obtain more regularized solutions and to show the existence of solutions for almost elastic bodies. "Almost elastic" implies that a viscous quantity dealing with the viscosity is chosen by a very small number, which enables viscoelastic bodies of Kelvin–Voigt type to get closer to elastic bodies.

One of the major concerns of dynamic contact problems is to show conservation of energy for the elastic case or energy balance for the viscoelastic case. This is an open question and in [Ahn 2007; 2008; 2012] it has been investigated theoretically and numerically. However, proving it for the general case may be a very difficult task. In rigid-body dynamic problems with frictionless impact, showing conservation of energy depends on the coefficient of restitution (COR). If COR $= 1$, that is, for the elastic case, energy conserves, but if $0 \leq \text{COR} < 1$, that is, for the inelastic case, energy decreases. Furthermore, considering COR for particles results in showing the uniqueness of solutions, which is stated at the end of Section 4.

This work is motivated by the three-dimensional dynamic contact problem, although particles are neither elastic nor viscoelastic. Our dynamic contact model may be basic but will provide a great opportunity to think about significant issues on higher-dimensional dynamic contact problems with elastic bodies or rigid bodies.
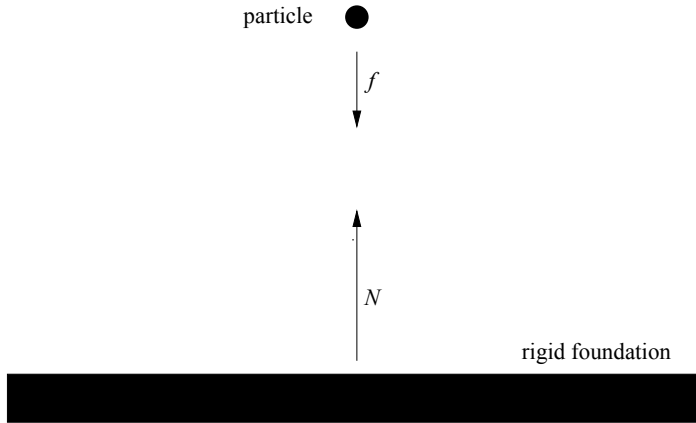
Figure 1. Dynamic contact of a particle.

2. Continuous formulations and some mathematical backgrounds

The motion of a particle in this physical situation is described by the ordinary differential equation (ODE)

$$u_{tt} = N + f \quad \text{for all } t \in (0, T],$$

where $u = u(t)$ is the displacement of a particle, $f = f(t)$ is a given body force, and $N = N(t)$ is a contact force. The acceleration of the particle, u_{tt}, is the second derivative of u with respect to time t, and T is the final time for the motion of the particle. When the particle drops down and hits the fixed flat rigid obstacle φ and bounces off, the Signorini contact conditions are applied which can be understood in terms of complementarity conditions (CCs). In general, the CCs $0 \leq a \perp b \geq 0$ mean that the scalars a and b are nonnegative and either a or b is zero. Now, we can see that the contact conditions satisfy the CCs (2-2) where the flat rigid foundation φ does not depend on time t. When there is a gap between the particle and the rigid foundation ($u(t) > \varphi$), the contact force N must be zero, and when the particle is in contact with the rigid foundation ($u(t) = \varphi$), that is, there is no gap, the contact force takes place ($N(t) \geq 0$). We note that $u(t) \geq \varphi$ implies that the particle does not penetrate the rigid foundation unless the normal compliance applies to the stationary foundation. By Newton's third law, the contact forces N are always regarded as nonnegative. The physical situation is illustrated in Figure 1.

Thus, we establish the ODE and the CCs that describe the physical situation: for all $t \in (0, T]$,

$$u_{tt}(t) = N(t) + f(t), \tag{2-1}$$

$$0 \leq u(t) - \varphi \perp N(t) \geq 0, \tag{2-2}$$

$$u^0 = u(0), \tag{2-3}$$

$$u_t^0 = u_t(0), \tag{2-4}$$

where u^0 is the initial displacement and u_t^0 is the initial velocity of the particle. For our convenience, it can be assumed that the flat rigid foundation $\varphi = 0$, without loss of generality. In order to prove the existence of solutions, (2-1) has to be considered in the sense of distributions and then we will seek solutions $u : [0, T] \to \mathbb{R}$ in appropriate spaces.

Let q and g be any functions. Then we introduce the little o notation:

$$q = o(g) \quad \text{provided} \quad \lim_{t \to \infty} \frac{|q(t)|}{|g(t)|} = 0.$$

This notation implies that the function g approaches infinity even faster than the function q does as $t \uparrow \infty$.

The Laplace transform of any function w, which is a useful tool for handling ODEs, is defined by

$$(\mathscr{L}w(t))(s) = \int_0^\infty w(t)e^{-st}\, dt. \tag{2-5}$$

It is important to take a restriction of the number s (possibly complex number) into consideration, in order to see the convergence of (2-5). Lemma 1 in Section 3 requires Lerch's theorem [Widder 1941, pp. 62–63]; generally speaking, it implies that if $(\mathscr{L}w)(s) = (\mathscr{L}\varpi)(s)$ with all s in some region of convergence, then $w(t) = \varpi(t)$ for *almost all $t \in [0, T]$*. This is called Lerch's cancellation law.

3. Conservation of energy

In this dynamic contact problem, the energy function $E(t)$ is defined by

$$E(t) := E[u, u_t] = \tfrac{1}{2}[u_t(t)]^2 - f(t)u(t), \tag{3-1}$$

where the first term and the second term in (3-1) are called the kinetic energy and the potential energy, respectively, and u_t denotes the velocity of a particle. One can see that the velocity u_t is replaced by the new variable v in Section 4.

If the conservation of energy is considered in terms of the *atom* level (see [Moreau et al. 1988]), its mathematical proof will be much harder. Showing conservation of energy might be the most difficult task in the dynamic contact problems with Signorini contact conditions. However, if functions are piecewise continuous, then the Laplace transform is one-to-one, which means that we can apply Lerch's cancellation law. In order to do so, we assume that the impact time period is not instantaneous; that is, the impact time period is $(t_* - \epsilon, t_* + \epsilon)$ with sufficiently small $\epsilon > 0$. In the following lemma, the minimum requirement is

that the displacement u is piecewise smooth which implies that u is differentiable almost everywhere and u_t has a jump discontinuity at a finite number of points.

Lemma 1. *Assume that there is no change of body force and the solutions u satisfying the continuous formulations (2-1)–(2-4) are piecewise smooth and $u(t) = \varphi$ for all $t \in (t_* - \epsilon, t_* + \epsilon)$ with the fixed $t_* \in (0, \infty)$. If $E = o(e^t)$ as $t \uparrow \infty$, then energy conserves; that is, $E(0) = E(t)$ for almost all $t \in (0, \infty)$.*

Proof. Multiplying both sides of (2-1) by the velocity u_t, we have $u_{tt}u_t - fu_t = Nu_t$. Since $(d/dt)(u_t^2/2) = u_{tt}u_t$, we can obtain

$$\frac{d}{dt}\left(\frac{u_t^2}{2} - fu\right) = Nu_t.$$

Recall the CCs $0 \leq u(t) - \varphi \perp N(t) \geq 0$ with $0 < t_* < t \leq T$. There are two cases; if $N(t) = 0$, then $N(t)u_t = 0$ over the interval $(0, T]$, and if $N(t) > 0$ over $(t_* - \epsilon, t_* + \epsilon)$ and $N(t) = 0$ outside of $(t_* - \epsilon, t_* + \epsilon)$, then $u(t) = \varphi$ over $(t_* - \epsilon, t_* + \epsilon)$ and thus $N(t)u_t = 0$ on $(0, T]$. So $E(0) = E(t)$ for $t \in (0, T]$. Note the velocity u_t is piecewise continuous.

Now, we take the Laplace transform of both sides to get

$$\int_0^\infty \left(\frac{u_t^2}{2} - fu\right)' e^{-st}\,dt = \int_0^\infty Nu_t e^{-st}\,dt. \qquad (3\text{-}2)$$

Here $'$ means the derivative with respect to time t. Integrating by parts we get

$$0 = \int_0^\infty \left(\frac{u_t^2}{2} - fu\right)' e^{-st}\,dt$$

$$= \left[\left(\tfrac{1}{2}u_t^2 - fu\right)e^{-st}\right]_0^\infty + s\int_0^\infty \left(\frac{u_t^2}{2} - fu\right)e^{-st}\,dt. \qquad (3\text{-}3)$$

Since $E = o(e^t)$ as $t \uparrow \infty$, there is a constant $M > 0$ such that

$$|E(t)| = \left|\frac{u_t^2}{2} - fu\right| \leq Me^t \quad \text{for some large } t > 0.$$

Since we require the convergence of the improper integral on the right side of (3-3), we need to impose the condition that $1 - s < 0$. Thus it follows from (3-3) that, for $s > 1$,

$$(\mathscr{L}E(t))(s) = (1/2s)\big(u_t^2(0) - 2fu(0)\big).$$

We can also see that $(\mathscr{L}E(0))(s) = (1/2s)(u_t^2(0) - 2fu(0))$ for $s > 0$. Thus, we note that the Laplace transform requires one-to-one mapping for only $s > 1$. Since $(\mathscr{L}E(t))(s) = (\mathscr{L}E(0))(s)$ for $s > 1$, $E(0) = E(t)$ for almost all $t \in (0, \infty)$, as required. □

Remarks 2. In Lemma 1, the displacement u may be semismooth (see its definition in [Facchinei and Pang 2003b, Section 7.4]), since we have the condition $u(t) = \varphi$ for all $t \in [t_* - \epsilon, t_* + \epsilon]$ with the fixed $t_* \in (0, \infty)$.

Unfortunately, the technique used in Lemma 1 does not work for the viscoelastic or elastic cases, since it is relatively more difficult to handle the elastic energy included in the energy function.

4. Numerical formulations and their convergence

In this section, we set up three numerical equations based on the continuous formulations (2-1)–(2-2), with (4-2) being an extra equation where we set the change in the displacement equal to the average velocity between the time steps. First, we partition the time interval $[0, T]$ such that

$$0 = t_0 < t_1 < t_2 < \cdots < t_l < \cdots < t_{n-1} < t_n = T,$$

where n is the number of time steps. The uniform time step $h = T/n$ is used and thus the size of the time step is $h = t_{l+1} - t_l$ and each discretized time is $t_l = lh$ for any integers $l \geq 0$. Then, the numerical approximations $u(t_l)$, $v(t_l)$ and $N(t_l)$ are denoted by u^l, v^l and N^l, respectively. Assume that there is no change of body force f. Using the implicit Euler method (sometimes referred to as the backwards Euler method) for the CCs, we are led to the following numerical formulations:

$$\frac{v^{l+1} - v^l}{h} = N^l + f, \tag{4-1}$$

$$\frac{u^{l+1} - u^l}{h} = \frac{v^{l+1} + v^l}{2}, \tag{4-2}$$

$$0 \leq u^{l+1} - \varphi \perp N^l \geq 0. \tag{4-3}$$

The solutions (u, v, N) of our contact problem (2-1)–(2-4) will be approximated by the numerical trajectories (u_h, v_h, N_h), which satisfy the numerical formulations (4-1)–(4-3); let $u_h(t)$ be a piecewise linear interpolant satisfying $u(t_l) = u^l$ and $u(t_{l+1}) = u^{l+1}$, and let $v_h(t)$ be a piecewise constant interpolant satisfying $v(t) = v^{l+1}$ for $t \in (t_l, t_{l+1}]$. We also set up the numerical approximation $N_h(t)$ of the contact forces, which is the step function; that is, $N(t) = N^l$ for $t \in [t_l, t_{l+1})$ and thus the approximation N_h has to be defined in the distributional sense to be

$$N_h(t) = h \sum_{l=0}^{n-1} \delta(t - (l+1)h) N^l, \tag{4-4}$$

where δ is the Dirac delta function. We also define the energy function for the

discrete case to be

$$E(t_l) := E^l = \tfrac{1}{2}(v^l)^2 - fu^l, \tag{4-5}$$

which plays a very important role in showing the boundedness of numerical solutions from the theoretical perspective and addressing the stability in the numerical perspective.

Thanks to our numerical scheme, the numerical formulations (4-1)–(4-3) confirm the regularity of numerical solutions (u_h, v_h, N_h) for any $h > 0$. Lemma 3 demonstrates a possibility of energy conservation and supports the regularity of solutions.

Lemma 3. *Suppose that our numerical solutions satisfy the numerical formulations* (4-1)–(4-3) *for any time step $h > 0$ and the body force f is given as a constant function. If the initial data u^0, v^0 are finite, we have the following estimates*:

$$\max_{0 \le l \le n} |v^l| \le \sqrt{\left(2E^0 + 2|f||u^0| + |f|T\right)\left(1 + |f|Te^{|f|T}\right)} < \infty,$$

$$\max_{0 \le l \le n} |u^l| \le |u^0| + \frac{T}{2}\sqrt{\left(2E^0 + 2|f||u^0| + |f|T\right)\left(1 + |f|Te^{|f|T}\right)} < \infty.$$

Proof. Using (4-1) and (4-2), for any $h > 0$ we have

$$\frac{(v^{l+1})^2 - (v^l)^2}{2h} = \frac{N^l(u^{l+1} - u^l)}{h} + \frac{f(u^{l+1} - u^l)}{h}.$$

It follows from the numerical CCs that

$$\frac{(v^{l+1})^2 - (v^l)^2}{2} - f(u^{l+1} - u^l) = N^l(u^{l+1} - u^l)$$
$$= N^l[u^{l+1} - \varphi - (u^l - \varphi)]$$
$$= -N^l(u^l - \varphi) \le 0. \tag{4-6}$$

Therefore, from (4-6),

$$E^{l+1} = \tfrac{1}{2}(v^{l+1})^2 - fu^{l+1} \le \tfrac{1}{2}(v^l)^2 - fu^l = E^l.$$

So repeating the inequality at each time step $t = t_l$, we can get $E^l \le E^0$ for any $l \ge 1$. Thus,

$$\tfrac{1}{2}(v^l)^2 \le E^0 + fu^l \le E^0 + |f||u^l| \le E^0 + |f|\left(|u^0| + \frac{1}{2}\int_0^{t_l} |v_h(\tau)|\,d\tau\right).$$

Note that

$$|u^l| \le |u^0| + \frac{1}{2}\int_0^{t_l} |v_h(\tau)|\,d\tau. \tag{4-7}$$

Since v_h is a constant interpolant, by Cauchy's inequality, we can set up

$$|v_h(t_l)|^2 \leq 2E^0 + 2|f|\left(|u^0| + \frac{1}{2}T + \frac{1}{2}\int_0^{t_l} |v_h(\tau)|^2\, d\tau\right).$$

Using Gronwall's inequality, we have

$$(v^l)^2 = |v_h(t_l)|^2 \leq \left(2E^0 + 2|f||u^0| + |f|T\right)\left(1 + |f|Te^{|f|T}\right) \quad \text{for any } l \geq 0. \quad (4\text{-}8)$$

It also follows from (4-7)–(4-8) that

$$|u^l| \leq |u^0| + \frac{T}{2}\sqrt{\left(2E^0 + 2|f||u^0| + |f|T\right)\left(1 + |f|Te^{|f|T}\right)},$$

as desired.   □

We note that the estimates in Lemma 3 can be obtained even if the body force f is not a constant function. Now, we introduce notations to see how to show the existence of solutions. If $u : [0, T] \to \mathbb{R}$ is continuous, then the p-th Hölder norm of u is defined by

$$\|u\|_{C^p[0,T]} = \sup_{t \in [0,T]} |u(t)| + \sup_{s \neq t \in [0,T]} \frac{|u(t) - u(s)|}{|t - s|^p}.$$

Considering Hölder spaces would be useful to show the compactness of continuous solutions for PDEs. Applying Lemma 3 to the construction of numerical solutions, we can see that $u_h \in C[0, T]$ and $v_h \in L^\infty[0, T]$ for any time step size $h > 0$. However, showing the boundedness of solutions is not enough to prove the existence of solutions. Thus, we need compactness to show that u_h converges strongly in $C[0, T]$ as $h \downarrow 0$. Now, we choose any s_1, s_2 such that $0 \leq s_1 < s_2 \leq T$, $|s_1 - s_2| < h$, $s_1 \in (t_{l-1}, t_l]$, and $s_2 \in (t_l, t_{l+1}]$. We can use Lemma 3 again to have

$$
\begin{aligned}
|u_h(s_2) - u_h(s_1)| &= |u_h(s_2) - u_h(s_1)|^p |u_h(s_2) - u_h(s_1)|^{1-p} \\
&\leq \frac{1}{2}\left(\int_{s_1}^{s_2} |v_h(t - h) + v_h(t)|\, dt\right)^p \left(|u_h(s_2)| + |u_h(s_1)|\right)^{1-p} \\
&\leq C|s_2 - s_1|^p.
\end{aligned}
$$

Consequently, we can see easily that the interpolant $u_h \in C^p[0, T]$ with exponent $0 < p \leq 1$. By the Arzelà–Ascoli theorem, $C^p[0, T]$ is compactly embedded in $C[0, T]$. Therefore, there is a subsequence of u_h (denoting this sequence by u_h), such that u_h converges strongly to u, that is, $u_h \to u$ in $C[0, T]$, as $h \downarrow 0$.

We regard the numerical contact force N_h as the Borel measure on the time interval $[0, T]$:

$$N_h([0, T]) = \int_{[0,T]} N_h(t)\, dt.$$

Using (4-4), we can show the boundedness of N_h easily. Recalling the numerical formulation (4-1), we have

$$\int_{[0,T]} N_h(t)\, dt = h \sum_{l=0}^{n-1} N^l = v^n - v^0. \tag{4-9}$$

Equation (4-9) does make sense from a physical point view, since the velocity v moves down initially, and thus $v^0 < 0$ and the particles bounce off, and thus their velocity $v^n > 0$. Therefore, for any $h > 0$ we have

$$\int_{[0,T]} N_h(t)\, dt \leq \sqrt{\left(2E^0 + 2|f||u^0| + |f|T\right)\left(1 + |f|T e^{|f|T}\right)} - v^0 < \infty.$$

Applying the Riesz representation theorem [Renardy and Rogers 1993, p. 199] and Alaoglu's theorem [ibid., p. 209], N_h has a subsequence that is weakly* convergent to N in the sense of measures as $h \downarrow 0$. We denote the subsequence by N_h. Thus, $N_h \rightharpoonup^* N$. Finally, we check if our solutions, which converged by numerical trajectories, satisfy the CCs (2-2). Since $u_h - \varphi \geq 0$ and $u_h \to u$ as $h \downarrow 0$, we have $u - \varphi \geq 0$. Since $N_h \geq 0$ and $N_h \rightharpoonup^* N$ as $h \downarrow 0$, we have $N \geq 0$. We claim that $N(u - \varphi) = 0$ in the weak sense. Taking the integral of $N_h(u_h - \varphi)$, we have

$$\int_0^T N_h(t)(u_h(t) - \varphi)\, dt = h \int_0^T \sum_{l=0}^{n-1} \delta(t - (l+1)h) N^l(u_h(t) - \varphi)\, dt$$

$$= h \int_0^T \sum_{l=0}^{n-1} N^l(u^{l+1} - \varphi)\, dt = 0. \tag{4-10}$$

We notice that (4-10) is identified by the numerical CCs (4-3). Finally, we claim that $\int_0^T N_h(t)(u_h(t) - \varphi)\, dt \to \int_0^T N(t)(u(t) - \varphi)\, dt$. Since $u_h \to u$ and $N \rightharpoonup^* N$ as $h \downarrow 0$,

$$\left| \int_0^T N_h(t)(u_h(t) - \varphi)\, dt - \int_0^T N(t)(u(t) - \varphi)\, dt \right|$$

$$\leq \int_0^T \left| N_h(t)(u_h(t) - \varphi) - N(t)(u(t) - \varphi) \right| dt$$

$$= \int_0^T \left| N_h(t)(u_h(t) - \varphi) - N_h(t)(u(t) - \varphi) + N_h(t)(u(t) - \varphi) - N(t)(u(t) - \varphi) \right| dt$$

$$\leq \int_0^T \left| N_h(t)(u_h(t) - u) \right| dt + \int_0^T \left| (N_h(t) - N(t))(u(t) - \varphi) \right| dt \to 0.$$

Therefore, by the squeeze theorem, we can obtain

$$0 = \int_0^T N_h(t)(u_h(t) - \varphi)\, dt \to \int_0^T N(t)(u(t) - \varphi)\, dt \quad \text{as } h \downarrow 0.$$

Thus, we conclude that there exist solutions $u \in C[0, T] \cap C^p[0, T] \cap W^{1,\infty}[0, T]$ with $0 < p \leq 1$ satisfying (2-1)–(2-4), where the space $W^{1,\infty}[0, T]$ is defined by $W^{1,\infty}[0, T] = \{u \mid \sup_{0 \leq t \leq T}(|u(t)| + |u_t(t)|) < \infty\}$. We notice that the derivative u_t has to be considered in the weak sense.

Lemma 4 requires an additional condition that the solutions are absolutely continuous. We denote by $COR(u)$ the coefficient of restitution for the particle which is defined by $COR(u) = -v_a/v_b$, where v_a is the velocity after contact and v_b is the velocity before contact. Therefore, solutions that we seek have to be considered in the stronger sense in order to prove their uniqueness. We note that showing the uniqueness is trivial unless we take contact forces into consideration.

Lemma 4. *Suppose that there exist two solutions (u, N_1) and (w, N_2) satisfying (2-1)–(2-4). If either the initial velocity $u_t^0 = 0$ and $u_t(t) = w_t(t) = 0$ for some $t \in [0, T]$ or $COR(u) = COR(w) = 1$, then the two solutions are the same; that is, $u(t) = w(t)$ for all $t \in [0, T]$ and $N_1(t) = N_2(t)$ for almost all $t \in [0, T]$.*

Proof. We assume that there exist two solutions (u, N_1) and (w, N_2) such that

$$u_{tt} = N_1(t) + f(t) \quad \text{and} \quad w_{tt} = N_2(t) + f(t). \tag{4-11}$$

Letting $z(t) = u(t) - w(t)$, it is easy to see that $z_{tt} = N_1(t) - N_2(t)$. Multiplying by z_t and taking the integral over $[0, t] \subset [0, T]$, we can obtain

$$\int_0^t z_{\tau\tau} z_\tau \, d\tau = \int_0^t (N_1(\tau) - N_2(\tau))(u_\tau(\tau) - w_\tau(\tau)) \, d\tau$$
$$= \int_0^t N_1(\tau)u_\tau(\tau) - N_1(\tau)w_\tau(\tau) - N_2(\tau)u_\tau(\tau) + N_2(\tau)w_\tau(\tau) \, d\tau.$$

In Lemma 1, it has been shown from the CCs (2-2) that

$$N_1(\tau)u_\tau(\tau) = N_2(\tau)w_\tau(\tau) = 0.$$

Using the two equations in (4-11) and applying integration by parts, we have

$$\frac{1}{2}\int_0^t \frac{d}{d\tau}(z_\tau^2(\tau)) \, d\tau$$
$$= -\int_0^t N_1(\tau)w_\tau(\tau) + N_2(\tau)u_\tau(\tau) \, d\tau$$
$$= -\int_0^t u_{\tau\tau}(\tau)w_\tau(\tau) + w_{\tau\tau}(\tau)u_\tau(\tau) \, d\tau$$
$$= -\left(u_t(t)w_t(t) - u_t(0)w_t(0) - \int_0^t u_\tau(\tau)w_{\tau\tau}(\tau) \, d\tau + \int_0^t w_{\tau\tau}(\tau)u_\tau(\tau) \, d\tau\right)$$
$$= -u_t(t)w_t(t) + u_t(0)w_t(0). \tag{4-12}$$

If the initial two velocities $u_t(0) = w_t(0) = 0$ and $u_t(t) = w_t(t) = 0$ for some $t \in [0, T]$, it is easy to see from (4-12) that

$$\frac{1}{2} \int_0^t \frac{d}{d\tau}(z_\tau^2(\tau)) \, d\tau = z_t^2(t) - z_t^2(0) = 0. \tag{4-13}$$

If $\mathrm{COR}(u) = \mathrm{COR}(w) = 1$, the identities (4-13) also hold. Since the two solutions satisfy the initial data (2-4), from both cases we have $z_t^2(0) = 0$, which gives us $u_t(t) = v_t(t)$ for almost all $t \in [0, T]$. Therefore, $u(t) = w(t)$ for all $t \in [0, T]$ and the corresponding contact forces $N_1(t) = N_2(t)$ for almost all $t \in [0, T]$, as required.   □

When we consider Equation (4-12), we could impose the more general condition that $\mathrm{COR}(u), \mathrm{COR}(w) \geq 1$. However, the condition requires that the obstacle is deformed. Therefore, the uniqueness is shown under the assumption that particles collide with the rigid foundation elastically.

5. Numerical results and discussion

In this section, numerical results are presented implementing several methods. For the sake of simplicity, we assume that $\varphi = 0$ throughout this section. Even though we use different methods with $f = 0$ and without considering the coefficient of restitution, we obtain almost equivalent numerical results (simulations) which are displayed in Figures 2–3. These results may enable us to demonstrate some evidence for the numerical stability. Lemma 1 is proven by the main idea that our numerical schemes guarantee that energy does not increase. We shall observe numerical results later on that show the numerical evidence for energy conservation. This means that the numerical solutions are stable, because they satisfy the criterion that solutions never show increasing energy. The first method that we describe is an infinite-dimensional approach that has a completely different perspective from the other two numerical schemes. Indeed, the infinite-dimensional approach is motivated by the normal compliance (see [Klarbring et al. 1988]). If the contact conditions are rather changed to the normal compliance condition, the contact forces N will be replaced by

$$N(t) = p(u(t) - \varphi),$$

where a prescribed function p can be defined by $p(r) = c_N \max(r, 0)$ for $c_N \geq 0$ and c_N is called the normal compliance stiffness coefficient. As we shall see in Lemma 5, contact forces satisfying Signorini contact conditions (or CCs) can be approximated as $c_N \uparrow \infty$. Instead of using Signorini contact conditions, the normal compliance condition enables us to consider well-conditioned dynamic contact problems and more realistic physical situations. In Section 5.1, we shall use the

normal compliance to see how to construct approximations, depending on the parameter of penetration, $\epsilon > 0$. Mathematically speaking, the normal compliance plays a fundamental role in showing better regularity of solutions and the uniqueness of solutions for dynamic contact problems. In Section 5.2, we shall discuss two numerical methods based on time discretization; one is directly implemented from the numerical CCs (4-3) and another is carried out with the nonsmooth Newton's method. There is a classification for dynamic contact problems on $\mathbb{R}^d$ with $d \geq 1$; one class is a class of thick obstacle problems and the other is a class of boundary thin obstacle problems. The meaning of "thick" is that obstacles (or constraints) are applied over a subset of the whole domain, while the meaning of "thin" is that the obstacles are applied on a subset of only the boundary of the domain. Readers who are interested in this classification may refer to [Ahn and Stewart 2006]. Concerning the corresponding numerical schemes for the two classes, the nonsmooth Newton's method will be very useful and efficient for thick obstacle problems and it is not necessary for the boundary thin obstacle problems in the case that $d = 1$.

5.1. *Numerical results via the infinite-dimensional approach.* Our physical interpretation is that particles touch and penetrate a rigid obstacle over the short contact time period $(t_* - \epsilon, t_* + \epsilon)$ with $\epsilon > 0$. We assume that the solutions to the ODE (2-1) are smooth enough. Then, assuming that $f(t) = 0$, we can construct the natural cubic splines to interpolate the solutions:

$$
\begin{cases}
S_1(t) = -\dfrac{3}{2}(t - t_* + \epsilon) + \dfrac{1}{2\epsilon^2}(t - t_* + \epsilon)^3 & \text{on } (t_* - \epsilon, t_*], \\
S_2(t) = -\epsilon + \dfrac{3}{2\epsilon}(t - t_*)^2 - \dfrac{1}{2\epsilon^2}(t - t_*)^3 & \text{on } [t_*, t_* + \epsilon),
\end{cases}
$$

where ϵ is called an approximate parameter of penetration. The piecewise linear functions below are included in the entire solution for the displacement:

$$
\begin{cases}
S_3(t) = -\frac{3}{2}(t - t_*) - \frac{3}{2}\epsilon & \text{on } [0, t_* - \epsilon], \\
S_4(t) = \frac{3}{2}(t - t_*) - \frac{3}{2}\epsilon & \text{on } [t_* + \epsilon, \infty).
\end{cases}
$$

So $S_3(t)$ and $S_4(t)$ are the outer functions for the piecewise solution for the displacement of the particle. Let $S_\epsilon = S_1 \cup S_2 \cup S_3 \cup S_4$ be an approximation of the solutions u. Then, this approximation S is a smooth function, but it does not satisfy Signorini contact conditions. In Lemma 5, we will see that the approximation of the contact forces N_ϵ satisfying the normal compliance condition converges to δ in the distributional sense, as $\epsilon \downarrow 0$. Since we expect contact forces over the interval $(t_* - \epsilon, t_* + \epsilon)$, we consider the translated Dirac delta function $\delta(t - t_*)$ in Lemma 5. Let Ω be a nonempty open set in $\mathbb{R}$. Then, the set of all test functions on Ω is denoted by $\mathcal{D}(\Omega)$.

Lemma 5. *Let $N_\epsilon = S_1'' \cup S_2''$ be an approximation of contact forces over the interval $(t_* - \epsilon, t_* + \epsilon)$ for $t_* > 0$ with small $\epsilon > 0$. Then, $N_\epsilon \to \delta$ in the sense of distributions.*

Proof. We consider the sequence of contact forces as follows:

$$N_\epsilon(t) := \frac{1}{6} \begin{cases} S_1''(t) = \dfrac{3}{\epsilon^2}(t - t_* + \epsilon) & \text{if } t \in (t_* - \epsilon, t_*], \\[2mm] S_2''(t) = \dfrac{3}{\epsilon} - \dfrac{3}{\epsilon^2}(t - t_*) & \text{if } t \in [t_*, t_* + \epsilon), \\[2mm] \qquad 0 & \text{if } t \in (0, t_* - \epsilon] \cup [t_* + \epsilon, \infty). \end{cases}$$

Then, we claim that, for any test function $\psi \in \mathscr{D}(\mathbb{R}^+)$,

$$\int_0^\infty N_\epsilon(t)\psi(t)\,dt \to \int_0^\infty \delta(t - t_*)\psi(t)\,dt \quad \text{as } \epsilon \downarrow 0.$$

For any fixed $t_* > 0$ and $\epsilon > 0$ we define the integral functions F_1 and F_2 to be

$$F_1(\tau) = \int_{t_*}^\tau (t - t_* + \epsilon)\psi(t)\,dt,$$

$$F_2(\tau) = \int_{t_*}^\tau \left(1 - \frac{1}{\epsilon}(t - t_*)\right)\psi(t)\,dt \quad \text{for } \tau > 0.$$

Thus, it follows that

$$\int_0^\infty N_\epsilon(t)\psi(t)\,dt$$

$$= \frac{1}{2\epsilon^2}\int_{t_*-\epsilon}^{t_*}(t - t_* + \epsilon)\psi(t)\,dt + \frac{1}{2\epsilon}\int_{t_*}^{t_*+\epsilon}\left(1 - \frac{1}{\epsilon}(t - t_*)\right)\psi(t)\,dt$$

$$= \frac{1}{2\epsilon}\frac{F_1(t_* - \epsilon) - F_1(t_*)}{-\epsilon} + \frac{1}{2}\frac{F_2(t_* + \epsilon) - F_2(t_*)}{\epsilon} = \frac{1}{2\epsilon}\frac{dF_1(t_*)}{dt} + \frac{1}{2}\frac{dF_2(t_*)}{dt}.$$

By the fundamental theorem of calculus, part 2, we can obtain

$$\int_0^\infty N_\epsilon(t)\psi(t)\,dt = \tfrac{1}{2}\psi(t_*) + \tfrac{1}{2}\psi(t_*) = \psi(t_*) = \int_0^\infty \delta(t - t_*)\psi(t)\,dt \quad \text{as } \epsilon \downarrow 0,$$

which implies that $N_\epsilon \to \delta$ in the distributional sense as $\epsilon \downarrow 0$, as desired. $\qquad\square$

The approximation S_ϵ computed with the small parameter $\epsilon = 10^{-3}$ is presented in Figure 2. The top of Figure 2 is a visual representation of the natural cubic splines ($S_1(t)$ and $S_2(t)$ and their applicable derivatives) for the displacement, velocity, and contact forces. We can observe a little penetration of a particle due to the parameter $\epsilon = 10^{-3}$. In addition, we can guess that the less the penetration depth is, the larger the magnitude of the contact forces is. While the cubic splines only consider the time period when the particle is in contact with the rigid obstacle, the piecewise linear functions $S_3(t)$ and $S_4(t)$ and their applicable derivatives are

(a) Displacement u

(d) Displacement u

(b) Velocity v

(e) Velocity v

(c) Contact force N

(f) Contact force N
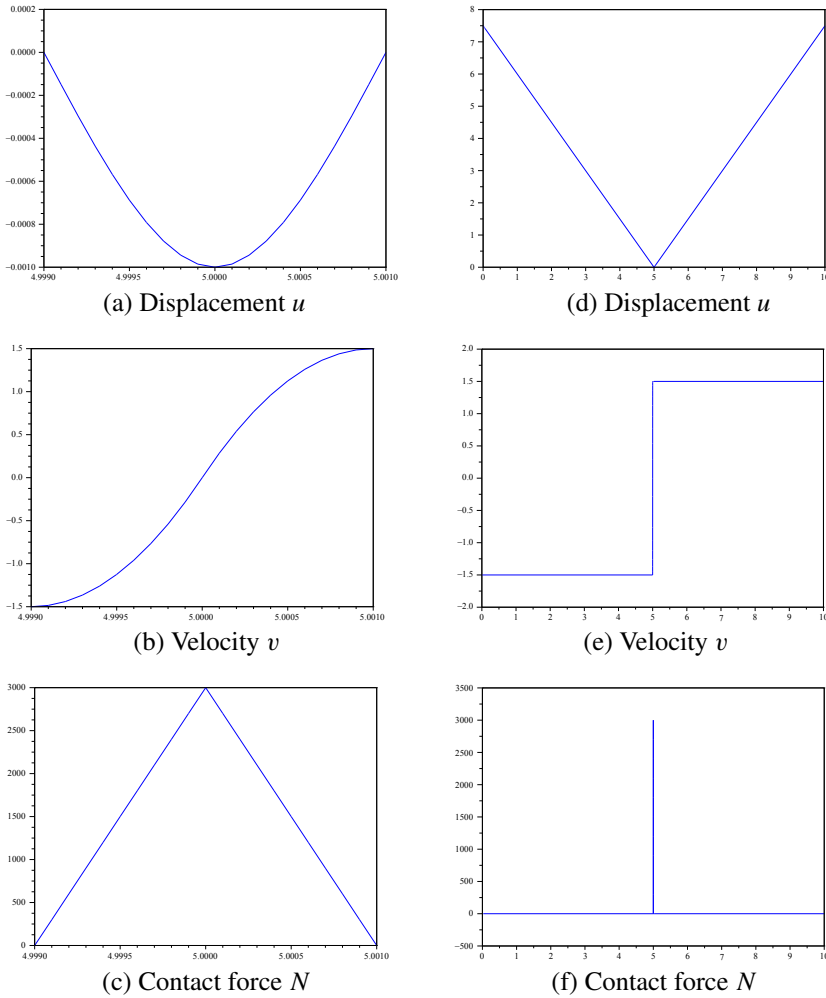
Figure 2. For $\epsilon = 0.001$ and $t_* = 5$, the graphs on the left, (a)–(c), represent the natural cubic splines for u, v, and N over the short time period $[t_* - \epsilon, t_* + \epsilon]$; the graphs on the right, (d)–(f), represent the entire piecewise functions for u, v, and N.

added to the ends of the splines to get the total picture of what is really happening throughout the particle's motion. This can be seen in the graphs in the right-hand column of Figure 2. Unfortunately, this infinite approach does not work in the dynamic adhesive contact model; see [Wolf 2012].

5.2. *Numerical results via the finite-dimensional approach.* In this subsection, two different numerical schemes are introduced and it is assumed that the body force f is a constant.

First, we explain our numerical scheme where we can directly compute the next step solution from the numerical CCs (4-3). The numerical equations (4-1)–(4-3) can be manipulated so that we obtain the solutions (u^{l+1}, N^l) at the next time step $t = t_{l+1}$. Using (4-2), from (4-1) we can solve for the next step solution u^{l+1}:

$$u^{l+1} = h\left(\frac{h(N^l + f)}{2} + v^l\right) + u^l. \tag{5-1}$$

The next step solution u^{l+1} needs to satisfy the CCs (4-3). So, if $u^{l+1} > \varphi$, we accept the solution (u^{l+1}, N^l) with $N^l = 0$. If $u^{l+1} = \varphi$, then we need to compute the previous contact force N^l:

$$N^l = \frac{2}{h}\left(\frac{\varphi - u^l}{h} - v^l\right) - f.$$

Once the next step solution u^{l+1} is obtained, we can compute the next step velocity v^{l+1} from the extra equation (4-2):

$$v^{l+1} = \frac{2}{h}(u^{l+1} - u^l) - v^l.$$

Secondly, we apply the nonsmooth Newton's method to compute u^{l+1}. Basically, solutions of dynamic contact problems are not smooth, because of the nature of the CCs. However, we can reformulate the approach by substituting a smooth function; see [Facchinei and Pang 2003a, p. 73 ff.]. One of the functions commonly used for this purpose is the Fischer–Burmeister function F, given by

$$F(a, b) = (a + b) - \sqrt{a^2 + b^2}. \tag{5-2}$$

It is not hard to see that $0 \le a \perp b \ge 0$ is equivalent to the equation $F(a, b) = 0$. This function is not still applied practically. In order to avoid the singularity happening, we set up the approximate function

$$F_\varepsilon(a, b) = (a + b) - \sqrt{a^2 + b^2 + \varepsilon}$$

for sufficiently small $\varepsilon > 0$, where ε is called a smoothing parameter. As $\varepsilon \downarrow 0$, $F_\varepsilon(a, b) \to F(a, b)$ in the strong sense.

Thanks to the equations (4-1)–(4-2), we can express the previous contact force N^l in terms of the next step solution u^{l+1}:

$$N^l = \frac{2}{h}\left(\frac{u^{l+1} - u^l}{h} - v^l\right) - f.$$

Thus, finding the next step solution u^{l+1} satisfying the CCs (4-3) is equivalent to finding the solution u^{l+1} satisfying the following nonlinear equation:

$$\left(u^{l+1} + \left[\frac{2}{h} \left(\frac{u^{l+1}}{h} - \frac{u^l}{h} - v^l \right) - f \right] \right)$$

$$= \sqrt{(u^{l+1})^2 + \left[\frac{2}{h} \left(\frac{u^{l+1}}{h} - \frac{u^l}{h} - v^l \right) - f \right]^2} + \varepsilon. \quad \text{(5-3)}$$

Now, we move the right side of (5-3) and replace the left side by the nonlinear function $S_\varepsilon(u^{l+1})$. So the next step solution u^{l+1} can be found for nonlinear equation $S_\varepsilon(u^{l+1}) = 0$. In order to compute the next step solution u^{l+1}, we can set up Newton's iterative formula:

$$u^{l+1}_{m+1} = u^{l+1}_m - \frac{S_\varepsilon(u^{l+1}_m)}{S'_\varepsilon(u^{l+1}_m)},$$

where u^{l+1}_{m+1} is the next solution and u^{l+1}_m is the previous solution for Newton's iteration. We note that S'_ε does not contain any singularity.

Based on the numerical equations (4-1)–(4-3), we tested the two numerical schemes. The results, which are almost indistinguishable, are shown in Figures 3 and 4, using an initial displacement of $u^0 = 5$, an initial velocity of $v^0 = -1$, an end time $T = 10$, and the step size $h = 0.001$. The body force f is not applied in this numerical experiment. When we implement the nonsmooth Newton's method, the smoothing parameter $\varepsilon = 10^{-15}$ is used and 10^{-15} is used for the stop criterion.

As can be seen in the left column of graphs in Figure 3, with no coefficient of restitution, the particle's motion reflects that of an absolute value function. Also note that we see a very similar graph as our natural cubic spline for the particle's displacement (as was displayed in Figure 2). Its velocity resembles the Heaviside function, as expected from our continuous result for the velocity of the particle. The impulse function δ can be seen in the graph of the contact force. The bottom left picture in Figure 3 supports conservation of energy numerically.

Numerically, we would also like to consider the particle's motion with a given coefficient of restitution since this would be more realistic. To change our numerical code to take the COR into account, we must alter the velocity at the instant that the particle is in contact with the surface where

$$0 \leq \text{COR} = \frac{-v_a}{v_b} \leq 1.$$

As expected, when a coefficient of restitution is introduced into our numerical formulations (in this inelastic case COR $= 0.75$), both the displacement and the velocity of the particle are dampened after impact; see Figure 3, right column. The implementation of a coefficient of restitution has no effect on the results of the contact force, but does have a rather large effect on the graph of the energy function. With a coefficient of restitution, we see that energy is lost after the particle's impact,
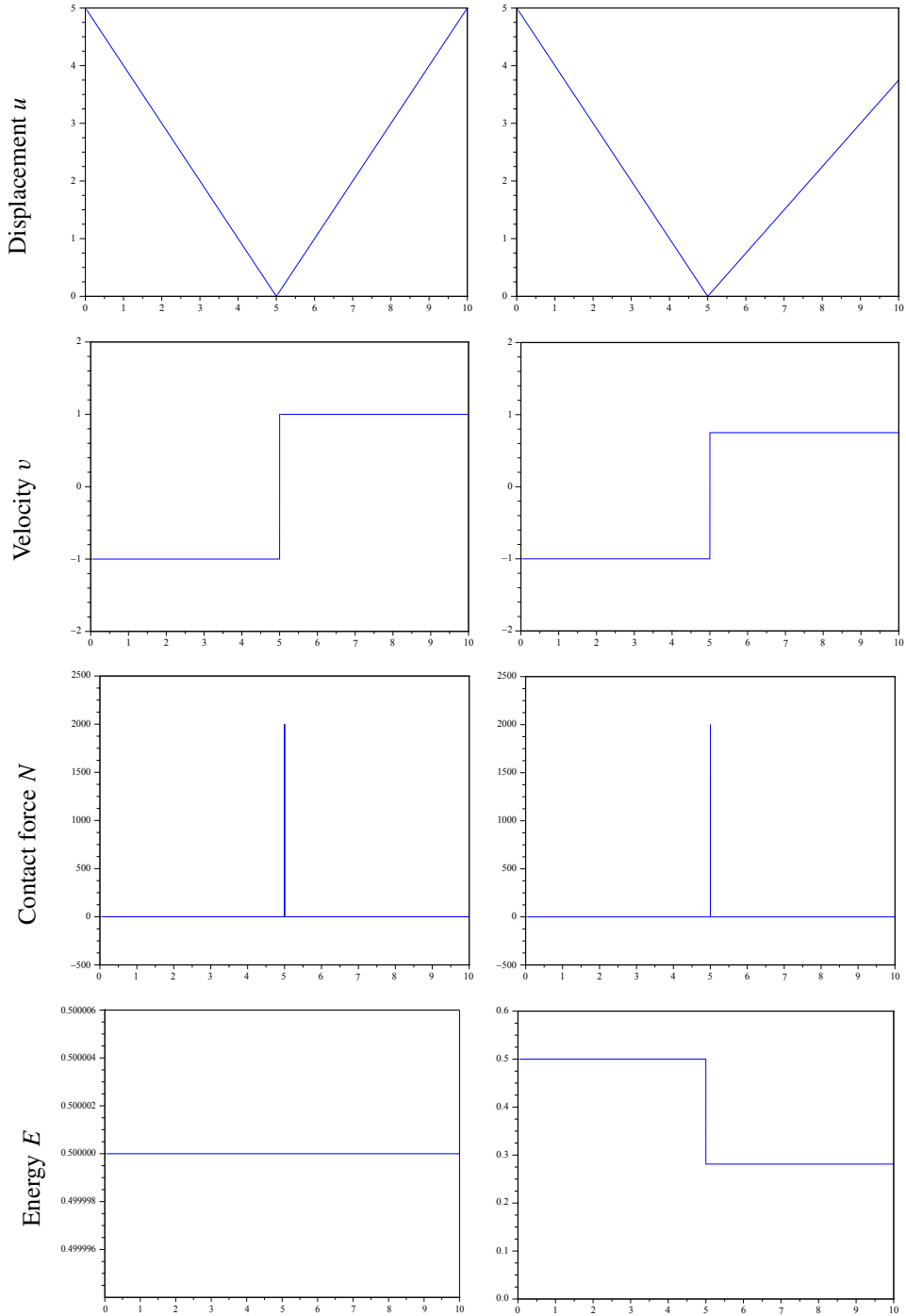
Figure 3. Numerical results without considering the coefficient of restitution (left) and with COR = 0.75 (right).

which can be shown theoretically. Also, the nonsmooth Newton's method with the Fischer–Burmeister equation still works very well when a coefficient of restitution is thrown into the mix.

Going back to our original numerical equation of motion (4-1), we note that we still need to incorporate a body force into the system. In a real-world sense of the situation, there is no better choice for a body force to impose on the particle than one that resembles Earth's gravitational force.

With this gravity-like body force, $f(t) = -9.80665$, we see some interesting graphs in Figure 4. The left column shows the simulations without a coefficient of resitution. The top graph, for the displacement, shows that the body force causes the particle to repeatedly bounce off the rigid obstacle until coming to a stop at around $t = 7.5$ for the selected initial conditions. However, we note that the height of the bounces does not decrease at a constant rate when only a body force is applied. The velocity shows a continual "zig-zag" centered about a velocity of zero. Conceptually, we can agree that the body force would continually pull the particle down, causing an increasingly negative velocity before bouncing back up, causing a jump of the function to a positive value, before falling again. Graphs of the contact forces each show multiple Dirac deltas, whose magnitude decreases over time until the particle comes to rest. With the energy function, like the contact and displacement graphs, energy decreases in steps with just a body force applied. Without considering a coefficient of restitution one might expect that energy conserves. Indeed, as the time step size h_t gets smaller and smaller, numerical simulations show that the energy function becomes flatter than the graph in the left column of Figure 4.

The application of both a coefficient of restitution and a body force combines to give us the most realistic solutions possible when thinking of a real-world situation. As seen in the right-hand column of Figure 4, the coefficient of restitution, in addition to the body force, gives us solutions for the displacement, velocity, and energy function that trend more steadily in comparison to those on the left column.

6. Conclusion

In this paper, we consider a second-order ODE with constraints. The existence of solutions is proved by using time discretization and passing to the limit as the time step size h decreases to zero. Although conservation of energy and uniqueness are proven in this paper under some restrictive assumptions, they are still open questions in general. Several numerical methods are introduced to present simulations which support conservation of energy. The two numerical methods provide almost identical results when we use the same input data. Therefore, our numerical schemes seem to be reasonably stable. In our future work, we will investigate a possibility of doing
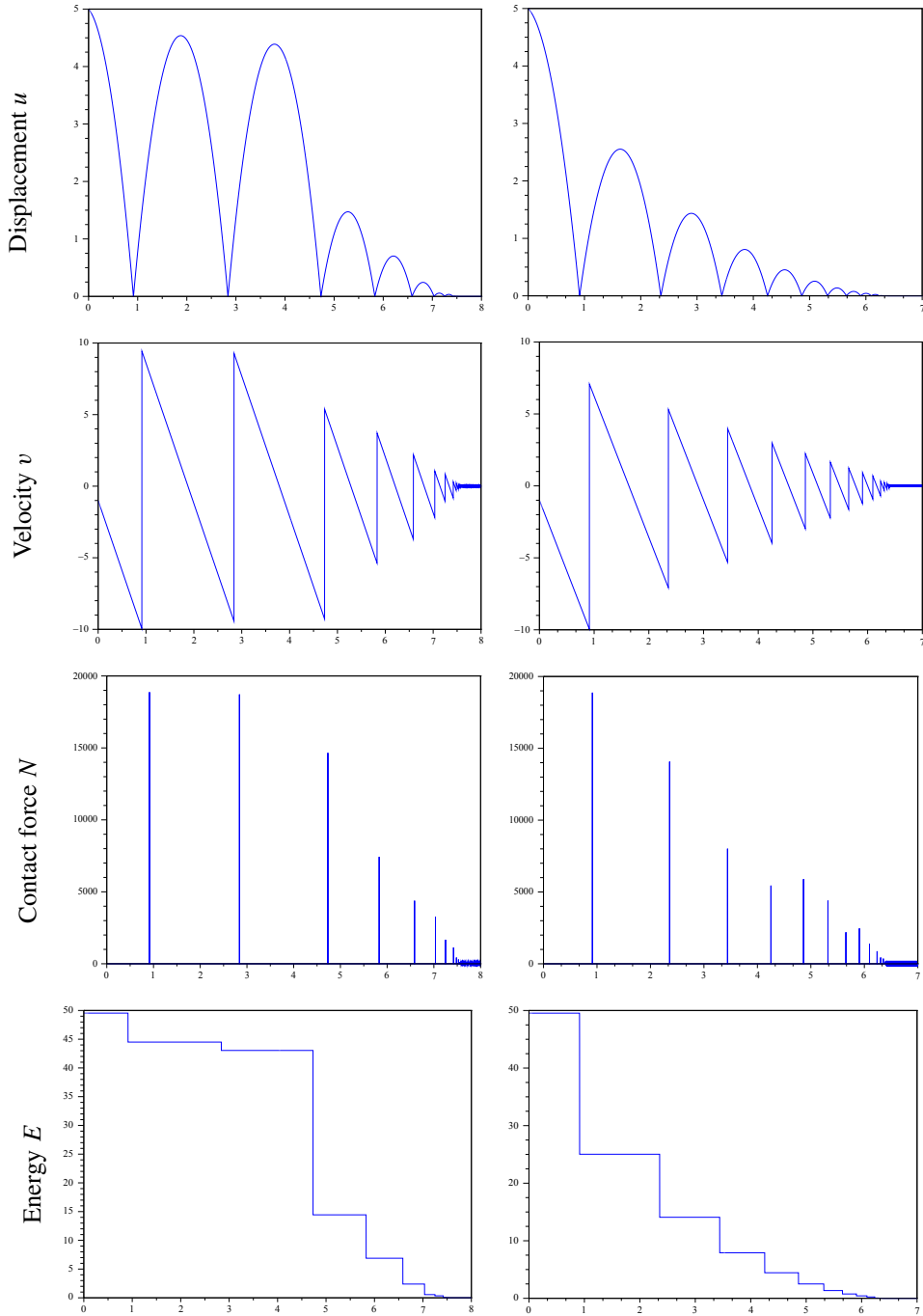
Figure 4. Numerical results with a body force of $f(t) = -9.80665$, representing gravity: without considering a coefficient of restitution (left column) and with COR = 0.75 (right column).

error analysis and study a more realistic contact model (see [Wolf 2012]) where we add the effect of a bonding field.

Acknowledgments

References

[Ahn 2007] J. Ahn, "A vibrating string with dynamic frictionless impact", *Appl. Numer. Math.* **57**:8 (2007), 861–884. MR 2008c:35166 Zbl 1114.74021

[Ahn 2008] J. Ahn, "Thick obstacle problems with dynamic adhesive contact", *M2AN Math. Model. Numer. Anal.* **42**:6 (2008), 1021–1045. MR 2009j:74072 Zbl 1149.74043

[Ahn 2012] J. Ahn, "A viscoelastic Timoshenko beam with Coulomb law of friction", *Appl. Math. Comput.* **218**:13 (2012), 7078–7099. MR 2880294 Zbl 06056829

[Ahn and Stewart 2006] J. Ahn and D. E. Stewart, "Existence of solutions for a class of impact problems without viscosity", *SIAM J. Math. Anal.* **38**:1 (2006), 37–63. MR 2007g:35158 Zbl 1116.35096

[Ahn and Stewart 2009] J. Ahn and D. E. Stewart, "Dynamic frictionless contact in linear viscoelasticity", *IMA J. Numer. Anal.* **29**:1 (2009), 43–71. MR 2010b:74018 Zbl 1155.74029

[Ahn et al. 2012] J. Ahn, K. L. Kuttler, and M. Shillor, "Dynamic contact of two Gao beams", *Electron. J. Diff. Equ.* **2012**:194 (2012), 1–42. MR 3001680

[Facchinei and Pang 2003a] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems, I*, Springer, New York, 2003. MR 2004g:90003a Zbl 1062.90001

[Facchinei and Pang 2003b] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems, II*, Springer, New York, 2003. MR 2004g:90003b Zbl 1062.90002

[Gao 1996] D. Y. Gao, "Nonlinear elastic beam theory with application in contact problems and variational approaches", *Mech. Res. Comm.* **23**:1 (1996), 11–17. MR 96m:73026 Zbl 0843.73042

[Hertz 1881] H. Hertz, "Über die Berührung fester elastischer Körper", *Journal für die reine und angewandte Mathematik* **92** (1881), 156–171. Zbl 14.0807.01

[Kikuchi and Oden 1988] N. Kikuchi and J. T. Oden, *Contact problems in elasticity: a study of variational inequalities and finite element methods*, SIAM Studies in Applied Mathematics **8**, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1988. MR 89j:73097 Zbl 0685.73002

[Klarbring et al. 1988] A. Klarbring, A. Mikelić, and M. Shillor, "Frictional contact problems with normal compliance", *Internat. J. Engrg. Sci.* **26**:8 (1988), 811–832. MR 89j:73098 Zbl 0662.73079

[Moreau et al. 1988] J.-J. Moreau, P. D. Panagiotopoulos, and G. Strang (editors), *Topics in nonsmooth mechanics*, Birkhäuser, Basel, 1988. MR 89c:00054 Zbl 0646.00014

[Renardy and Rogers 1993] M. Renardy and R. C. Rogers, *An introduction to partial differential equations*, Texts in Applied Mathematics **13**, Springer, New York, 1993. MR 94c:35001 Zbl 0917.35001

[Signorini 1933] A. Signorini, "Sopra alcune questioni di elastostatica", *Atti della Societa Italiana per il Progresso delle Scienze* (1933).

[Stewart 2000] D. E. Stewart, "Rigid-body dynamics with friction and impact", *SIAM Rev.* **42**:1 (2000), 3–39. MR 2001c:70017 Zbl 0962.70010

[Widder 1941] D. V. Widder, *The Laplace transform*, Princeton Mathematical Series **6**, Princeton University Press, 1941. MR 3,232d Zbl 0063.08245

[Wolf 2012] J. R. Wolf, *Dynamic contact of a particle: Mathematical theories and numerical approaches*, honors thesis, Arkansas State University, 2012, http://dbellis.library.astate.edu/vwebv/holdingsInfo?bibId=1541047.

jahn@astate.edu   *Department of Mathematics and Statistics,*
                                Arkansas State University, P.O. Box 70,
                                State University, AR 72467, United States

jared.wolf@smail.astate.edu   *Department of Mathematics and Statistics,*
                                Arkansas State University, P.O. Box 70,
                                State University, AR 72467, United States

Magic polygrams

Amanda Bienz, Karen A. Yokley and Crista Arangala

(Communicated by Filip Saidak)

Magic polygrams, which are extensions of magic squares, can be found with computer programs through exhaustive searches. However, most polygrams are too large for this method. Thus, these possibilities must be limited algorithmically. This paper investigates both a large traditional hexagram and a traditional octagram. Systematic approaches based on the arrangement of even and odd numbers are used to identify solutions.

1. Introduction

Magic squares are arrangements of numbers in which the orientation of the numbers leads to particular properties. Across the world, people have regarded the construction of magic squares as a form of mathematical study or a form of artistic creation, and the squares themselves were often believed to be objects with inherent good or evil powers [Cammann 1960].

Definition 1. In a *magic square*, nonnegative numbers are arranged so that all rows, columns, and main diagonals all sum to the same number. The common sum is known as the *magic constant*. A *traditional magic square* is an $n \times n$ magic square that is filled with the numbers 1 to n^2 [Beck et al. 2003; Benjamin and Yasuda 1999; Xin 2008].

The first magic square is thought to have originated from the *Lo Shu* diagram in the 23rd century BC, which is an orientation of dots supposedly originally revealed on the shell of a sacred turtle [Cammann 1961]. Although the legend is probably a more recent fabrication, this 3×3 traditional magic square was considered by ancient Chinese as a deeply meaningful symbol [Biggs 1979; Cammann 1961]. Scholars in the Middle East and India also studied magic squares, placing greater importance on them than their current classification in fields such as recreational mathematics [Cammann 1969a; Datta and Singh 1992]. Methods of construction of magic squares have varied greatly by culture and often reflected the philosophies

8	1	6
3	5	7
4	9	2

traditional

$k+a$	$k+a-b$	$k+b$
$k-a+b$	k	$k+a-b$
$k-b$	$k+a+b$	$k-a$

generic form
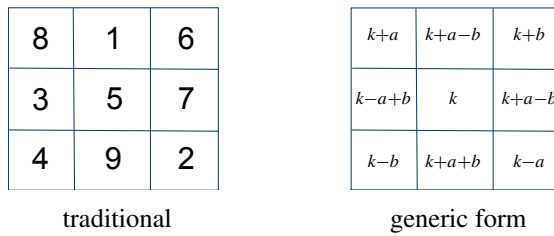
Figure 1. A filled-in 3×3 magic square and a generic version that works for any k, a, b [Chernick 1938].

of the particular culture [Biggs 1979; Cammann 1969a; 1969b; Datta and Singh 1992]. In India, 4×4 squares have been worn as amulets to bring luck [Datta and Singh 1992]. To some who studied the mysticism of magic squares, the knowledge that no 2×2 magic square exists was thought to reflect the imperfection of the four elements taken alone [Calder 1949]. As a result of the significance placed on magic squares, solutions have long been identified for traditional magic squares for $n = 3, 4, 5, 6, 7, 8, 9, 10$ [Biggs 1979; Cammann 1960].

Figure 1, left, shows a 3×3 traditional magic square with a magic constant of 15. Using the variables a, b, and k, every 3×3 magic square can be represented by a single pattern, as shown in the right half of the figure. There is only one unique solution to the traditional 3×3 magic square [Chernick 1938]. Increasing from a 3×3 magic square to a 4×4 magic square increases both the number of possible arrangements and the number of ordinary solutions. There are 880 unique 4×4 traditional magic squares [Beck et al. 2003].

A spin-off of the magic square is the magic hexagram.

Definition 2. A *hexagram* is a star with six points containing an arrangement of twelve numbers (see Figure 2).

Unlike a magic square, a magic hexagram is considered to contain only "rows". As seen in Figure 2, a hexagram contains six rows of five triangles each, and a total of twelve triangles. If you use the numbers 1 through 12 to fill every triangle in the
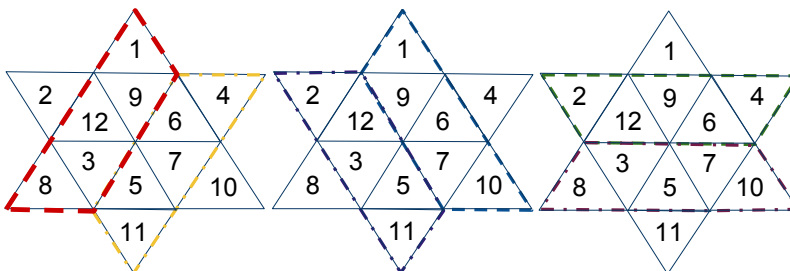
Figure 2. Examples of hexagrams with outlines of every row.

hexagram there are 12! different, but not necessarily unique, arrangements. While a computer can use brute force to find solutions for magic squares, attempting to identify solutions for a hexagram with 12! possible arrangements is computationally challenging. The number of possible arrangements can be reduced algorithmically in order to find solutions; see [Bolt et al. 1991; Gardner 2000].

Due to symmetry, hexagrams that are reflections or rotations of one another are equivalent. Eliminating these equivalent hexagrams, the number of possible unique arrangements is reduced to 11!; see [Bolt et al. 1991; Gardner 2000]. To eliminate the remaining duplicates, complementary solutions can be ignored [Gardner 2000].

Definition 3. [Gardner 2000] A *complementary arrangement* is obtained by subtracting each number of a polygram from the pattern's largest number plus 1.

Using Definition 3, the complement to a 3×3 square can be found be subtracting every number from 10. While the complement of a magic square is just a rotation of the original magic square, this is not the case for many other shapes [Gardner 2000]. To further reduce possibilities, the arrangements of even and odd numbers can be examined. An odd/even arrangement is a pattern of zeros and ones in which the ones represent odd numbers and the zeros represent even numbers [Bolt et al. 1991; Gardner 2000]. Because all rows of a magic hexagram must have a common sum, there must be either an odd number of odds in every row or an even number of odds in every row, limiting possible odd/even patterns. Ignoring transformations, there are only six different ways that even and odd numbers can be arranged throughout the hexagram [Gardner 2000]. However, odd/even patterns that represent complementary solutions are trivial variations of one another and can be ignored, limiting the number of odd/even patterns.

The patterns of many complementary polygrams are opposites. Odd/even patterns of a magic hexagram are considered complements when every number in one pattern is the opposite in the other pattern. While these patterns are different, there exists a complementary magic polygram for every magic polygram, so they are considered trivial variations [Gardner 2000].

In the current paper, two polygrams are investigated to find magic arrangements. Both the magic extended hexagram and the magic octagram are polygrams for which a complete list of solutions is difficult to identify. In order to identify these magic polygrams, odd/even arrangements are investigated, and upper and lower bounds of the magic constant are considered as in [Gardner 2000]. To limit the number of arrangements, distinct odd/even patterns are found in which every row has an equivalent number of odds. Methods such as the investigation of odd/even diagrams and magic constant bounds are used to focus the investigation because the number of total arrangements in the polygrams makes computational investigations challenging. The number of arrangements is further limited by finding the possible
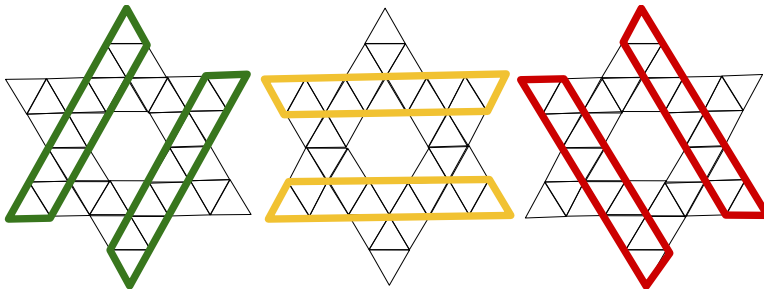
Figure 3. Extended hexagram structure with outlines of every row.

magic constants for each odd/even pattern. The goal of this research is to identify solutions to magic extended hexagrams and to magic octagrams, but not necessarily to find an exhaustive list. Only traditional polygrams are considered.

2. Generalizations about polygrams

The two polygrams that are considered in this study are (1) an extension of the hexagram described in Section 1 [Gardner 2000] and (2) another arrangement we will refer to as an octagram. By taking the magic hexagram described in Section 1 and outlining it, a larger hexagram with more triangles can be formed.

Definition 4. An *extended hexagram* is an arrangement of numbers in the shape of a six pointed star composed of 42 total triangles as shown in Figure 3.

The extended hexagram contains six rows of eleven triangles each. Figure 3 highlights the six different rows of this hexagram.

A similar shape to the hexagram can be formed using an octagon as the center of the diagram instead of a hexagon.

Definition 5. An *octagram* is a star with eight points containing an array of 16 numbers. An octagram has eight rows, containing six numbers each (see Figure 4).

In order to find magic extended hexagrams and magic octagrams, general results that apply to multiple polygrams can be identified. For instance, the odd/even pattern and its complement can be either the same or exact opposites depending on the number of positions in the shape.
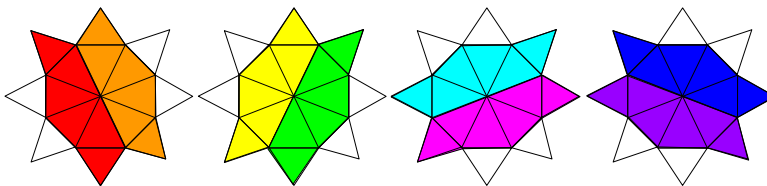
Figure 4. Octagram structure with outlines of the eight rows.

Theorem 1. *If a traditional magic polygram has an odd number of integers arranged within it, then the odd/even pattern of this polygram and its complement are equivalent.*

Proof. Suppose a magic polygram has $2N + 1$ numbers arranged throughout it, where N is some integer. Let $O_1, O_2, \ldots, O_{2N+1}$ represent the numbers in the original polygram. Let $C_1, C_2, \ldots, C_{2N+1}$ represent the numbers in the equivalent positions of the complementary polygram. By Definition 3, for some position P, $C_P = ((2N + 1) + 1) - O_P$, or $(2N + 2) - O_P$. There are two cases for what O_P could be. If O_P is even, $O_P = 2K$ for some integer K. So, $C_P = (2N + 2) - (2K)$, or $2(N - K + 1)$. This represents an even number, so when O_P is even C_P is also even. For the second case, when O_P is an odd number, it can be represented by $2K + 1$. So, $C_P = (2N + 2) - (2K + 1)$, or $2(N - K) + 1$. This represents an odd number, so when O_P is odd, C_P is also odd. As a result, the complementary odd/even pattern is the same as the odd/even pattern for any traditional polygram with an odd number of integers arranged throughout it. $\square$

Theorem 2. *If a traditional magic polygram has an even number of integers arranged within it, then the odd/even pattern of this polygram and its complementary pattern are opposites.*

Proof. Suppose a magic polygram has $2N$ numbers arranged throughout it, where N is some integer. Let $O_1, O_2, \ldots, O_{2N}$ represent the numbers in the original polygram. Let $C_1, C_2, \ldots, C_{2N}$ represent the numbers in the equivalent positions of the complementary polygram. By Definition 3, $C_P = (2N + 1) - O_P$ for some position P. There are two cases for what O_P could be. If O_P is even, $O_P = 2K$ for some integer K. So, $C_P = (2N + 1) - (2K)$, or $2(N - K) + 1$. This represents an odd number, so when O_P is even, C_P is odd. For the second case, when O_P is an odd number, it can be represented by $2K + 1$. So, $C_P = (2N + 1) - (2K + 1)$, or $2(N - K)$. This represents an even number, so when O_P is odd, C_P is even. As a result, the complementary odd/even pattern is the opposite of the original odd/even pattern for any traditional polygram with an even number of integers arranged throughout it. $\square$

If the number of odds per row of a traditional polygram is fixed, the number of odds in certain positions can be calculated.

Theorem 3. *If a magic polygram containing O odd numbers has N rows, each with K odds, and if a polygram is made up of only $\mathcal{X}$ positions appearing in A rows and $\mathcal{Y}$ positions appearing in B rows, then there are $(BO - NK)/(B - A)$ odds in $\mathcal{X}$ positions and $(NK - AO)/(B - A)$ odds in $\mathcal{Y}$ positions.*

Proof. Suppose a magic polygram has O odd numbers arranged throughout N different rows, and that each row contains K odds. Also, suppose the polygram

is made up of $\mathscr{X}$ positions which hold X odds and appear in A rows, as well as $\mathscr{Y}$ positions which hold Y odds and appear in B rows. Because every $\mathscr{X}$ position appears in A rows and every $\mathscr{Y}$ position appears in B rows, $AX + BY$ represents the number of times an odd is a part of a sum of a row. Also, because K is the number of times an odd is part of a sum of each row, and there are N rows, then NK is the total number of times an odd is part of a sum of a row in the total sum. Hence,

$$NK = AX + BY. \tag{1}$$

Every number is placed in either an $\mathscr{X}$ position or a $\mathscr{Y}$ position, so the sum of the number of odds in $\mathscr{X}$ positions and the number of odds in $\mathscr{Y}$ positions is equivalent to the number of odds placed in this hexagram. Hence,

$$X + Y = O. \tag{2}$$

By combining (1) and (2),

$$Y = (NK - AO)/(B - A).$$

By replacing Ys with $(NK - AO)/(B - A)$, Equation (2) becomes

$$X = (BO - NK)/(B - A).$$

Therefore, to have K odd numbers as part of a sum of each row, there must be $(NK - AO)/(B - A)$ odd numbers in $\mathscr{X}$ positions and $(BO - NK)/(B - A)$ odd numbers in $\mathscr{Y}$ positions. $\qquad\square$

The lower boundary of a magic constant can be found by placing the smallest numbers in the positions that appear in the largest number of rows. The upper boundary magic constant is found by doing the opposite. A magic polygram containing N rows is made up of only $\mathscr{X}$ positions and $\mathscr{Y}$ positions such that there are M number of $\mathscr{X}$ positions appearing in A rows, and O number of $\mathscr{Y}$ positions appearing in B rows, and the number contained in the i-th $\mathscr{X}$ position is X_i and the number in the i-th $\mathscr{Y}$ position is Y_i. The magic constant of a magic polygram is equivalent to the total sum of all of the rows combined, divided by the number of rows, so the magic constant is equal to

$$\frac{1}{N}\left(A \sum_{i=1}^{M} X_i + B \sum_{i=1}^{O} Y_i \right). \tag{3}$$

3. Traditional magic extended hexagram

If the numbers 1 through 42 are placed throughout the extended hexagram (as shown in Figure 3), there are 42! (total) arrangements possible. In order to reduce

this number and make identification of magic traditional extended hexagrams computationally easier, the number of possibilities is reduced by considering odd/even diagrams and by only considering nontrivial variations of a particular scenario.

Throughout the extended hexagram, 24 of the triangles (or numerical positions) appear in two different rows, while the remaining 18 triangles only appear in one row each. There are two ways for an extended hexagram to be magic. The first is that every row has an odd number of odd numbers, and then the magic constant is an odd number. The second is that every row has an even number of odd numbers, resulting in an even magic constant. One specific scenario is when every row has the same number of odd numbers, and the current paper will focus on this particular case.

3.1. *The possible number of odds in a row.* To find the possible arrangements of even and odd numbers throughout the hexagram where every row has the same number of odds, the first step is finding the overall number of odd numbers in all of the rows combined. Each number in the extended hexagram can be considered as in an $\mathscr{X}$ position (appearing in two rows) or in a $\mathscr{Y}$ position (appearing in one row). Because there are a total of twenty-one odd numbers placed throughout the six rows in the extended hexagram, O is 21 and N equals 6 in Theorem 3. $\mathscr{X}$ positions appear in two rows and $\mathscr{Y}$ positions appear in one, so A is 2 and B is 1. Using Theorem 3, the number of odds in $\mathscr{X}$ positions can be found by

$$X = 6K - 21 \tag{4}$$

and the number of odds in $\mathscr{Y}$ positions is

$$Y = 42 - 6K. \tag{5}$$

To find all distinct odd/even patterns in which every row has the same number of odds, the possibilities of numbers of odds per row should first be found. $\mathscr{X}$ positions appear in more rows than $\mathscr{Y}$ positions, so the maximum number of odds per row can be found by maximizing the number of odds in $\mathscr{X}$ positions. Because there are twenty-one odds placed throughout $\mathscr{X}$ and $\mathscr{Y}$ positions, when the number of odds in $\mathscr{X}$ positions is maximized, the number in $\mathscr{Y}$ positions is minimized. Because there are twenty-one odds and twenty-four $\mathscr{X}$ positions, the smallest possible number of odds in $\mathscr{Y}$ positions is zero.

By minimizing the number of odds in $\mathscr{Y}$ positions, (5) is set equal to zero. Solving $42 - 6K = 0$ for K shows that $K = 7$ when there are no odds in $\mathscr{Y}$ positions. Replacing K with 7 in (4) and (5) results in $X = 21$ and $Y = 0$. Hence, when there are exactly seven odds per row, twenty-one odd numbers are in $\mathscr{X}$ positions and zero odds are in $\mathscr{Y}$ positions.

To find the minimum possible number of odds per row, the number of odds in $\mathcal{X}$ positions must be minimized. Because there are twenty-one odds and only eighteen $\mathcal{Y}$ positions, the minimum possible number of odds in $\mathcal{X}$ positions is three. Three odds in $\mathcal{X}$ positions ($6K - 21 = 3$) results in $K = 4$. Plugging this into (4) and (5) shows that $X = 3$ and $Y = 18$. So, for there to be exactly four odds in each row, there must be three odds in $\mathcal{X}$ positions and eighteen odds in $\mathcal{Y}$ positions.

Knowing that the number of odds per row must be between four and seven, the possibilities in which K is either five or six must also be investigated. Using (4) and (5), the number of odds in $\mathcal{X}$ and $\mathcal{Y}$ positions can be found. For there to be five odd numbers in each row, there must be nine odds in $\mathcal{X}$ positions and twelve odds in $\mathcal{Y}$ positions. To have exactly six odds per row, there must be fifteen odds in $\mathcal{X}$ positions and six in $\mathcal{Y}$ positions.

3.2. *Odd/even patterns.* As determined in the previous section, there are four different cases in which every row of a magic hexagram can have the same number of odds. These cases are

(Case 1) 7 odds per row; 21 odds in $\mathcal{X}$ positions and 0 odds in $\mathcal{Y}$ positions.

(Case 2) 6 odds per row; 15 odds in $\mathcal{X}$ positions and 6 odds in $\mathcal{Y}$ positions.

(Case 3) 5 odds per row; 9 odds in $\mathcal{X}$ positions and 12 odds in $\mathcal{Y}$ positions.

(Case 4) 4 odds per row; 3 odds in $\mathcal{X}$ positions and 18 odds in $\mathcal{Y}$ positions.

The patterns resulting from each of these cases can be complementary to each other, because the number of odds in $\mathcal{X}$ positions in one pattern is equivalent to the number of evens in $\mathcal{X}$ positions in the complementary pattern. The number of odds in $\mathcal{Y}$ positions in one pattern is also equivalent to the number of evens in $\mathcal{Y}$ positions in the other pattern. For any pattern in which the Case 1 holds, if all of the evens and odds are switched, Case 4 results. As a result, these two cases are complementary and one of them can be considered trivial. Case 2 and Case 3 are also complementary to each other. By considering Cases 3 and 4 as trivial, there are only two nontrivial cases in which all rows have the same number of odds: when there are either seven odds in each row (Case 1), or six odds in each row (Case 2).

There are many different ways to arrange patterns such that they have the same magic constant and are therefore trivial variations. Numbers that appear in only one row can be switched to other positions only appearing in the same row without changing the magic constant. Arrangements that only switch ones and zeros throughout these positions have equivalent odds in each row. Additionally, there exist pairs of positions that appear in only the same two rows. If the numbers are switched between these two positions, both numbers are still in the same two rows and hence contribute to the same row sum. These switches, as a result, are trivial in both the traditional extended magic hexagram and in the odd/even diagrams.
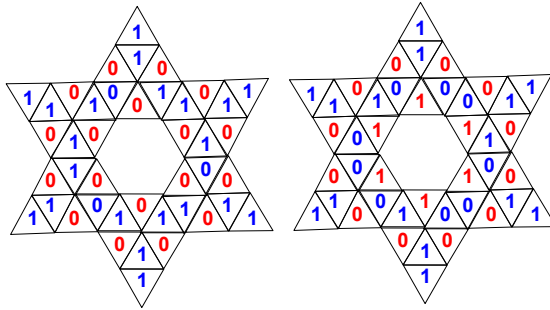
Figure 5. The two distinct odd/even arrangements for which every row in the extended hexagram has the same number of evens and odds.

Considering patterns defined for Case 1 and Case 2 and ignoring trivial variations, there are only two distinct arrangements for magic extended hexagrams for which every row has the same number of odd numbers. These two patterns are shown in Figure 5.

3.3. *Magic constants*. The upper and lower bounds of the magic constant can be found for both of these extended hexagrams using (3). Because the extended magic hexagram is made up of six rows, has twenty-four $\mathcal{X}$ positions each appearing in two rows, and has eighteen $\mathcal{Y}$ positions each appearing in only one row, (3) can be rewritten as

$$\frac{1}{6}\left(2\sum_{i=1}^{24} X_i + \sum_{i=1}^{18} Y_i\right). \tag{6}$$

Because $\mathcal{X}$ positions appear in a larger number of rows than $\mathcal{Y}$ positions, the lower bound magic constant is found by placing the lowest possible numbers in the $\mathcal{X}$ positions and the largest numbers in $\mathcal{Y}$ positions. The upper bound magic constant is found by doing the opposite.

For the hexagram with seven odd numbers in each row, all twenty-one odd numbers, along with three even numbers, are placed in positions that appear in two different rows. By placing the lowest possible numbers in the $\mathcal{X}$ positions and the largest in $\mathcal{Y}$ positions, (6) results in a lower bound of 226. By placing the largest numbers in $\mathcal{X}$ positions and smallest in $\mathcal{Y}$ positions, (6) shows the upper bound magic constant is 244.

Similarly, the lower and upper bound magic constants can be found for the hexagram with six odd numbers in each row. This hexagram contains fifteen odds in $\mathcal{X}$ positions and six odds in $\mathcal{Y}$ positions. Equation (6) shows this hexagram has an upper bound magic constant of 269 and a lower bound magic constant of 203.

The possible magic constants for each of the hexagrams can be further limited by taking into account whether the magic constant must be odd or even. The magic constant for the hexagram with seven odds in each row must be a number between 226 and 244. However, because the hexagram has exactly seven odds in each row, the magic constant, or sum of the row, must be an odd number. So, the magic constant for the hexagram with seven odds in each row must be an odd number between 227 and 243.

The magic constant for the hexagram with six odds in each row is a number between 203 and 269. However, if there are exactly six odds in every row, the magic constant must be an even number. So, the magic constant for this hexagram must be an even number between 204 and 268.

Using these limitations, a simple computer program was written in an attempt to find solutions. Initially, each even and odd position was labeled sequentially. The program initially placed each odd number in a position labeled odd, so that 1 was placed in the first odd position, 3 in the second odd position, and so on. Similarly, each even was placed in the appropriate even position. Each time the program looped through, either two of the even numbers or two of the odd numbers were swapped so that every combination of the numbers could be investigated in an attempt to find arrangements where every row summed up to the magic constant. Every magic constant was to be investigated, printing all solutions to a file. However, computer programs written in both Prolog and C were not able to finish running within weeks. A parallelized genetic algorithm was also not able to find any solutions when run for an extended period of time. While the algorithm greatly limited the possible arrangements of this extended hexagram, further reduction is necessary to systematically identify solutions.

3.4. *Solutions for magic hexagrams.* Magic extended hexagram solutions can be found for upper and lower bound magic constants by strategically placing numbers in the polygram. The upper bounds of the magic constant for the seven odds per row and six odds per row cases were found by placing the highest numbers in $\mathscr{X}$ positions, but placing the very highest numbers in these locations would not result in a solution. Similarly, lower bounds were identified for the magic constant in both cases by placing the lowest numbers in $\mathscr{X}$ positions, but this arrangement of the lowest numbers would not result in a solution. However, placing most of the high numbers or low numbers in $\mathscr{X}$ positions can lead to solutions through the recognition of patterns in the odd/even diagrams.

Figure 5 contains distinct odd/even arrangements for the seven odds per row and six odds per row cases; numbers in $\mathscr{X}$ positions are blue and numbers in $\mathscr{Y}$ positions are presented in red. In both diagrams, all $\mathscr{X}$ positions appear in diamond pairs, and each number in the pair affects the same rows in the hexagram. The diamonds can

be grouped into three categories:

- Pairs of odd numbers at the points of the star, or corner pairs, with sum C,

- Mixed pairs of one even and one odd in the internal hexagon, or mixed internal pairs, with sum M, or

- Pairs of two odds (7 odds per row) or two evens (6 odds per row) in the internal hexagon, or consistent internal pairs, with sum I.

In order to reduce the magic extended hexagram to a problem that is more easily solvable, the sums of the pairs in each category are set to be equal. Under these conditions, the magic constant will be equal to $2C + M + I$ plus the sum of the three numbers in the $\mathcal{Y}$ positions in any particular row. Additionally, by placing the highest or lowest values in particular positions, particular solutions can be identified.

As described in Section 3.3, the upper bound for the magic constant is 243 for the seven odds per row case. The upper bound was identified by placing all twenty-one odds in $\mathcal{X}$ positions and the highest evens in the remaining $\mathcal{X}$ positions. In order to identify magic extended hexagram solutions, the odd numbers from 19 to 41 are placed in corner pair positions such that each pair totals 60. To try to find a solution with a magic constant of 243, the highest evens should be placed in $\mathcal{X}$ positions if possible. Because the corner pairs now contain odds, high evens are placed in the only remaining $\mathcal{X}$ positions: the mixed internal pairs. If consecutive evens are chosen for these three locations, any list of three consecutive odds may be chosen in order to have a consistent sum M. As previously shown in Section 3.3, placing the three highest even numbers (38, 40, and 42) in the mixed internal pair positions will not result in a solution (because the magic constant must be odd in this case). However, placing the consecutive even numbers 36, 38, and 40 in mixed internal pair positions will lead to a solution (although using the numbers 34, 36, and 38 will not). The details of solutions of this form are as follows:

- The odd numbers 19 through 41 are placed in corner pair positions.

- The even numbers 36, 38, and 40 are paired with three consecutive odds (not already in corner pair positions) for the mixed internal pairs.

- The six remaining odds can be ordered $O_1 > O_2 > O_3 > O_4 > O_5 > O_6$ and then grouped into pairs (O_1 and O_6, O_2 and O_5, O_3 and O_4) for the consistent internal pairs.

- The 18 remaining evens are separated into groups of three, each with sum 58, and placed in the $\mathcal{Y}$ positions of six rows.

An example of a magic extended hexagram solution with this structure is shown in Figure 6, left.

Solutions for the magic extended hexagram can also be found by placing low numbers in $\mathcal{X}$ positions. The lower bound for the magic constant for the seven odds

per row case was found by placing all twenty-one odd numbers in $\mathcal{X}$ positions and the three lowest even numbers in the remaining $\mathcal{X}$ positions. In order to identify solutions with a magic constant of 227, odd numbers from 1 to 21 were placed in corner pair positions such that each pair totals 24. In order to identify a solution with a magic constant of 227, low even numbers were placed in mixed internal pair positions. The structure of existing examples of extended hexagrams with a magic constant of 227 can be identified using a similar method to that used to identify solutions for extended hexagrams with a magic constant of 243. The details of this solution arrangement are as follows:

- The odd numbers 1 through 21 are placed in corner pair positions.

- The even numbers 4, 6, and 8 are paired with three consecutive odds (not already in corner pair positions) for the mixed internal pairs.

- The six remaining odds can be ordered $O_1 > O_2 > O_3 > O_4 > O_5 > O_6$ and then grouped into pairs (O_1 and O_6, O_2 and O_5, O_3 and O_4) for the consistent internal pairs.

- The 18 remaining evens are separated into groups of 3, each with sum 74, and placed in the $\mathcal{Y}$ positions of six rows.

An example of a magic extended hexagram solution with this structure is shown in Figure 6, right. Complements for the magic extended hexagram solutions in Figure 6 are shown in Figure 7.

As shown in Figure 6, solutions to the magic extended hexagram exist for the seven odds per row case. The numbers can be moved around the hexagram and still result in a solution as long as the numbers stay in the same category pair or particular

magic constant = 243   magic constant = 227

Figure 6. Examples of magic extended hexagram solutions for the seven odds per row case when the magic constant is 243 (left) and 227 (right). Corner pairs are outlined in blue, mixed internal pairs in green, and consistent internal pairs in red.

magic constant = 230   magic constant = 246
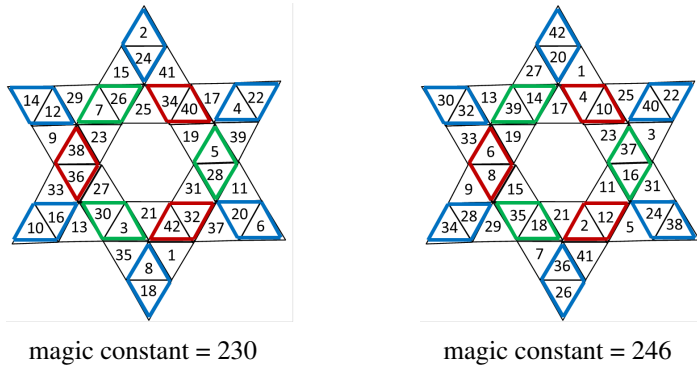
Figure 7. Complements of the solutions shown in Figure 6 (color conventions are the same).

set of three evens. Because these changes do not represent a simple rotation of the entire polygram, movement of numbers within category pairs or particular sets results in a new, nontrivial solution. Additionally, different lists of consecutive odds could have been chosen for the mixed internal pairs. Because all odd numbers are placed in one of the pairs (corner, mixed internal, or consistent internal), the odd numbers likely could have been arranged differently to also arrive at a solution. The complements shown in Figure 7 show examples of solutions for the four odds per row case and also reflect the structure of keeping the paired sums equal.

Magic extended hexagram solutions can also be found for the six odds per row case with upper and lower bound magic constants. In order to identify solutions that have a magic constant of 268, the largest odds are placed in corner pair positions. Because the largest even numbers can be placed in either of the internal pair blocks, two different methods are used to find solutions. In the first scenario, the six largest even numbers are placed in consistent internal pairs of the extended hexagram. More details of the construction of this solution are as follows:

- The odd numbers 19 through 41 are placed in corner pair positions.

- The even numbers 32 through 42 are placed in consistent internal pair positions.

- The next highest even numbers (26, 28, 30) are placed in mixed internal pair positions.

- All possible sets of three consecutive (remaining) odd numbers are investigated as potential numbers for the mixed internal pair positions. The only possible list is 11, 13, and 15, which forces the remaining three numbers in each row to total 33.

- The 18 remaining numbers are separated into groups of three (one odd and two evens), each with sum 33, and placed in the $\mathcal{Y}$ positions of six rows.
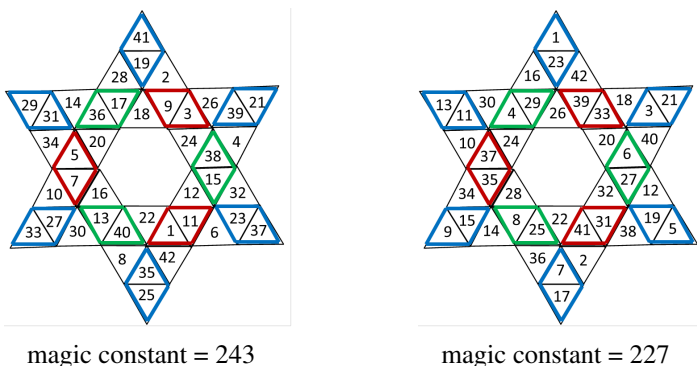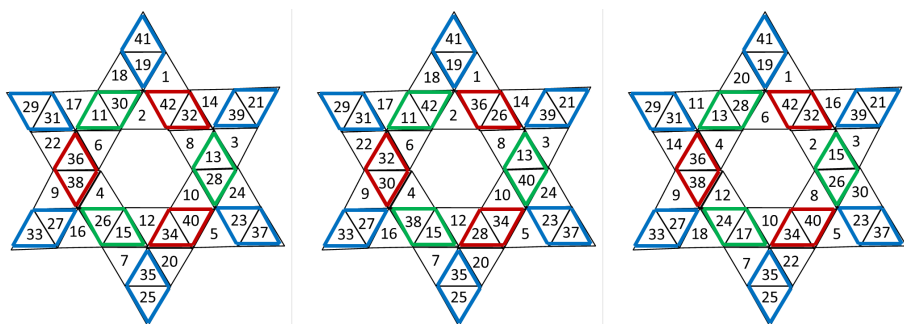
Figure 8. Examples of magic extended hexagram solutions for the six odds per row case when the magic constant is 268. Color conventions are as in Figure 6.

An example of a magic extended hexagram solution with this structure is shown in Figure 8, left. Alternatively, the three highest even numbers are placed in mixed internal pairs of the extended hexagram. More details of solutions of this form are as follows:

- The odd numbers 19 through 41 are placed in corner pair positions.
- The highest even numbers (38, 40, 42) are placed in mixed internal pair positions.
- The even numbers 26 through 36 are placed in consistent internal pair positions.
- All possible sets of three consecutive (remaining) odd numbers are investigated as potential numbers for the mixed internal pair positions. The only possible list is 11, 13, and 15, which forces the remaining three numbers in each row to total 33.
- The 18 remaining numbers are separated into groups of three (one odd and two evens), each with sum 33, and placed in the $\mathcal{Y}$ positions of six rows.

An example of a magic extended hexagram solution with this structure is shown in Figure 8, middle.

Note that the only differences in the solutions presented in the first two parts of Figure 8 are the locations of even numbers in $\mathcal{X}$ positions in the interior of the hexagram. However, this change is not the result of a rotation of the entire polygram or of the interior of the hexagram.

Other solutions to the magic extended hexagram with six odds per row can be found in similar ways. The more generalized process is this:

- The highest (or lowest) odd numbers are placed in corner pair positions.
- Even numbers are selected for the consistent internal pair positions. A set of three consecutive evens are placed in mixed internal pair positions.
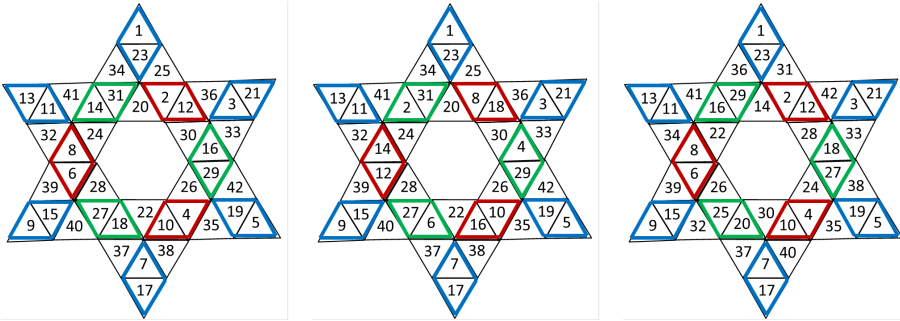
Figure 9. Examples of magic extended hexagram solutions for the six odds per row case when the magic constant is 204.

- All possible sets of three consecutive (remaining) odd numbers are investigated as potential numbers for the mixed internal pair positions. If a viable list is identified, the remaining 18 numbers are separated into groups of three each (one odd and two evens) and placed in the $\mathcal{Y}$ positions.

An additional example of a magic extended hexagram with magic constant 268 is presented in Figure 8, right. Examples of solutions with six odds per row when the magic constant is 204 are shown in Figure 9. Figures 10 and 11 contain complementary solutions to those in Figures 8 and 9, respectively.

As with the solutions presented for the seven odds per row and four odds per row cases, the numbers in the presented solutions for the six odds per row case can be moved around the hexagram within the same category pair and result in a solution. Not only can numbers be rotated within category pairs and result in a different solution, multiple structures are identified for the six odds per row case as shown in Figures 8 and 9. The complements to the six odds per row case represent
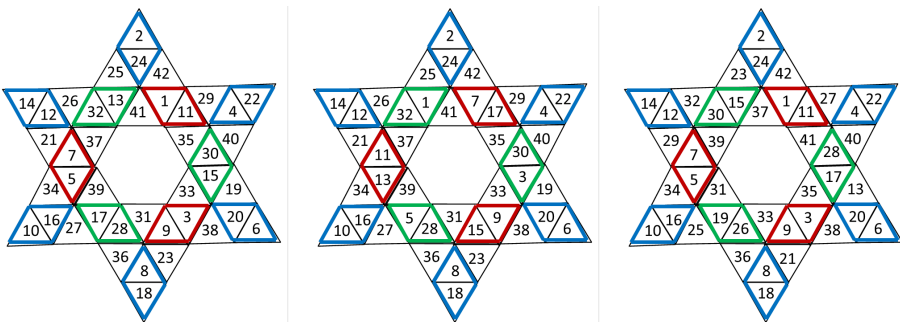
Figure 10. Complements of the magic extended hexagram solutions for the six odds per row case shown in Figure 8. The magic constant is 205.
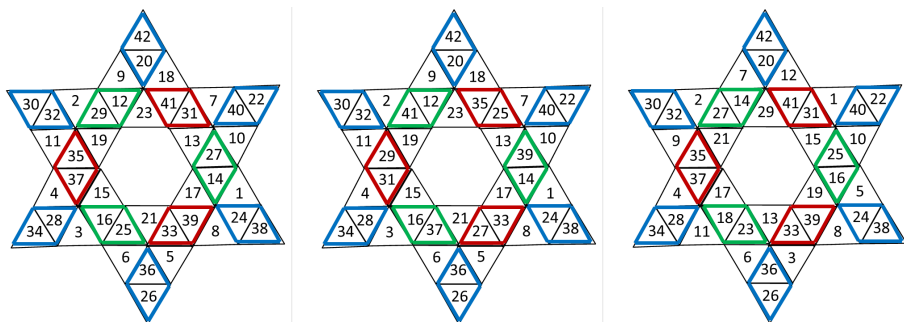
Figure 11. Complements of the magic extended hexagram solutions for the six odds per row case shown in Figure 9. The magic constant is 269.
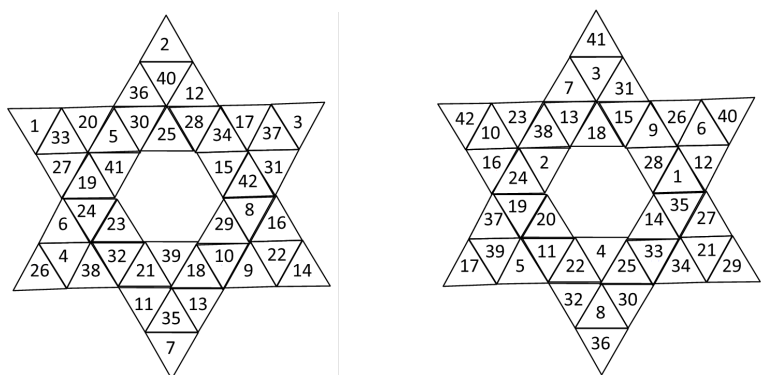
Figure 12. Example of a magic extended hexagram solution when the number of odds per row is not fixed (left), together with its complement (right).

solutions for the five odds per row case and also reflect the same structure. While many solutions have been identified for the cases when there are the same number of odds per row, a definitive list has not been established. Further, solutions for the magic extended hexagram do exist for cases when the number of odds per row is not fixed (as is shown in Figure 12).

4. Traditional magic octagram

The methods used to find magic extended hexagrams can also be applied to find other magic polygrams, such as magic octagrams. If numbers 1 through 16 are placed throughout the octagram (as shown in Figure 4), there are 16! (total) arrangements possible. In order to reduce this number as in the investigation on magic extended

hexagrams, the number of possibilities will be reduced by considering odd/even diagrams and by only considering nontrivial variations.

The inner positions of the octagram each appear in four different rows, while the outer positions each appear in two different rows. Similar to the magic hexagram, there are two possible cases in which every row can add up to a single number. The first case is when every row contains an odd number of odds. The magic constant, or common sum of each row, must be an odd number. The second case is when every row contains an even number of odds. In this case, the magic constant would be an even number. As in Section 3, this paper investigates only scenarios in which every row has exactly the same number of odds.

4.1. *Odd/even patterns.* As shown in Figure 4, the numbers in the eight positions within the central octagon appear in four rows each (which we will refer to as $\mathscr{X}$ positions), and the numbers in the eight positions that are the points of the star only appear in two rows each (which we will refer to as $\mathscr{Y}$ positions). Because a magic traditional octagram contains eight odd numbers and has eight rows with eight $\mathscr{X}$ positions and eight $\mathscr{Y}$ positions, Theorem 3 shows that there are

$$4K - 8 \tag{7}$$

odds in $\mathscr{X}$ positions and

$$16 - 4K \tag{8}$$

odds in $\mathscr{Y}$ positions.

Setting (8) to zero and solving for K shows that the maximum number of odds per row is four. Plugging this into K for (7) and (8) shows that for there to be exactly four odds in each row, there must be eight odd numbers in $\mathscr{X}$ positions and zero odds in $\mathscr{Y}$ positions. Setting (7) to zero and solving for K shows that the minimum number of odds per row is two. Equations (7) and (8) show that for K, or the number of odds per row, to be equal to two, there must be zero odds in $\mathscr{X}$ positions and eight odd numbers in $\mathscr{Y}$ positions.

Because the maximum possible number of odds per row is four and the minimum is two, the only other possible number of odd numbers per row is three. Substituting 3 for K in (7) and (8) shows that for there to be exactly three odds per row, there must be four odd numbers in $\mathscr{X}$ positions and four odds in $\mathscr{Y}$ positions.

The octagram investigated in this paper has an even number of integers arranged throughout it, so by Theorem 2, two odd/even patterns of this octagram are complements when the patterns are exact opposites of each other. The odd/even arrangement for when there are four odds in a row (all odds in the central octagon) is complementary to the odd/even arrangement when there are two odds in a row. In the process of finding magic octagrams, only one of these two patterns needs to be investigated. Figure 13 shows four of the distinct odd/even patterns for the

(a) 4 odds per row   (b) 3 odds per row   (c) 3 odds per row   (d) 3 odds per row
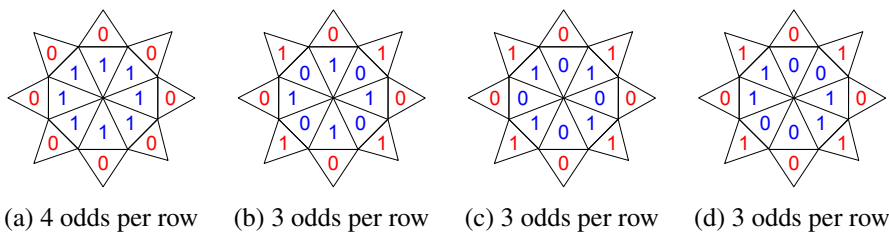
Figure 13. Four distinct odd/even patterns for the octagram.

magic octagram. There are multiple different patterns with three odds in every row, and all of the possible patterns for this scenario have not been investigated.

4.2. *Magic constants.* The upper and lower bounds of the magic constant can be found for distinct octagrams using (3). For the octagram, (3) can be rewritten as

$$\frac{1}{8}\left(4\sum_{i=1}^{8} X_i + 2\sum_{i=1}^{8} Y_i\right). \tag{9}$$

Similar to the process used to find the lower bound magic constant for hexagrams, the lower bound magic constant for each odd/even pattern of this octagram is found by placing the lowest possible numbers in $\mathcal{X}$ positions and the largest numbers in $\mathcal{Y}$ positions, and the upper bound magic constant is found by doing the opposite. For the octagram with four odd numbers in each row, all eight odd numbers are placed in the $\mathcal{X}$ positions and all eight even numbers are placed in the $\mathcal{Y}$ positions. Because there are only eight odd numbers to be placed in the eight $\mathcal{X}$ positions, there is only one possible magic constant rather than a range of possibilities. Using (9), the only possible magic constant for the octagram in which there are four odds per row is 50.

The lower and upper bound magic constants can be found for the octagrams with three odd numbers in each row. This octagram contains four odds in $\mathcal{X}$ positions and four odds in $\mathcal{Y}$ positions. Equation (9) shows that this octagram has an upper bound magic constant of 59 and a lower bound magic constant of 43.

4.3. *Solutions for magic octagrams.* By limiting the possible arrangements of numbers 1 through 16 throughout the octagram, the number of possibilities is small enough for a computer program to find magic octagrams. A brute force computer program with restrictions added in was written in C. Similar to the program described in Section 3, this program labeled all of the even and odd positions, initially arranged the numbers throughout their positions, and then stepped through every appropriate arrangement to find solutions. The program investigated each magic constant for every odd/even pattern, and with just these limitations, solutions were found in adequate time. There were hundreds of solutions for each of the four odd/even patterns in Figure 13.
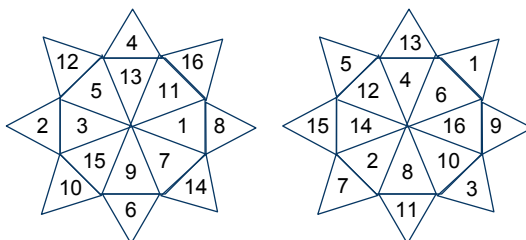
Figure 14. A magic octagram with four odds in each row (left) and a magic octagram with two odds in each row (right).

For the pattern in which all eight odd numbers are in $\mathscr{X}$ positions, as shown in Figure 13, there is only one possible magic constant, 50. For this magic constant, the 1920 different solutions were found computationally. The complementary pattern, in which all eight odd numbers are in $\mathscr{Y}$ positions, has 1920 solutions. Each of the solutions to the pattern with all eight odd numbers in $\mathscr{Y}$ positions is a complement to a solution of the pattern with all eight odd numbers in $\mathscr{X}$ positions.

The data for the other three patterns in Figure 13, each of which have three odds as a part of the sum in each row, is shown in Table 1. Pattern 1 has a total of 736 unique solutions, pattern 2 has 832, and pattern 3 has 1161, all among nine different magic constants. One solution for each pattern is shown on the top row of Figure 15. Each of these three patterns also has a complementary pattern. For every complementary pattern, there is an equivalent number of solutions, each one being the complement of an original solution. The complementary solutions for those shown in on the top row of Figure 15 are shown on the bottom row.

magic constant	pattern 1 Figure 13(b)	pattern 2 Figure 13(c)	pattern 3 Figure 13(d)
43	3	12	43
45	41	42	36
47	83	96	145
49	144	160	199
51	196	224	358
53	144	160	199
55	83	96	145
57	41	42	36
59	4	12	43

Table 1. Number of distinct octagram patterns of the forms shown in Figure 13(b)–(d).
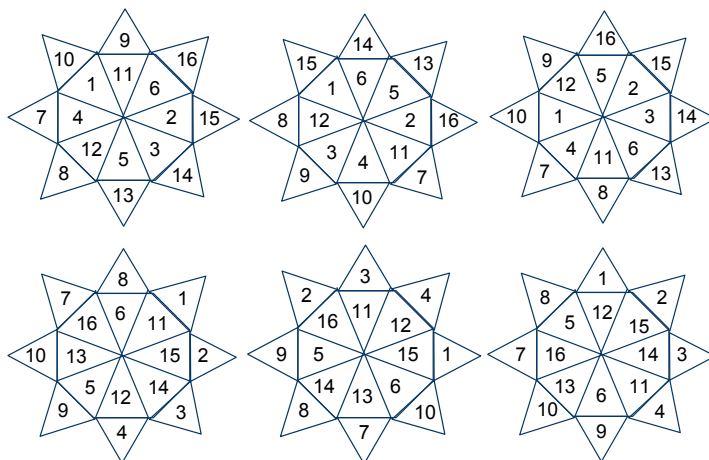
Figure 15. Top: solutions with a magic constant of 45 and three odds per row. Bottom: their complementary solutions, with magic constant 57.

5. Discussion and conclusions

If restrictions are placed on the number of even and odd numbers per row of the extended hexagram, characteristics of solutions can be identified. When there are seven odds per row, the magic constant of this extended hexagram must be an odd number between 227 and 243. The magic constant of the extended hexagram with six odds per row must be an even number between 204 and 268. Although an exhaustive list of solutions has not been established, multiple solutions have been identified.

Using similar restrictions on the number of evens and odds in the rows of the traditional extended hexagram, magic octagrams can also be identified. The possible magic constants for restricted patterns were found to further limit possibilities before using computer programs to identify specific solutions. The magic constant for the octagram with four odd numbers per row must be 50. The octagrams with three odds per row must have a magic constant between 43 and 59. The computationally discovered solutions were categorized by pattern and magic constant, but not all magic octagram solutions have necessarily been identified.

The only odd/even patterns that have been investigated for both polygrams are cases in which there are an equivalent number of odds in each row. The solutions for both polygrams when there are different numbers of odds in each row have not been investigated. Similarly, not every odd/even pattern in which there are three odds in every row of the octagram has been investigated. Additional studies could investigate more focused algorithms to find solutions in the fixed number of odds per row cases. Most hexagram solutions found in this study had the same number

of odds per row, but solutions do exist without this restriction as shown in Figure 12. Future investigations could focus on values in internal locations of the hexagram or octagram without the overall expectation of a certain number of odds per row in order to find solutions.

References

[Beck et al. 2003] M. Beck, M. Cohen, J. Cuomo, and P. Gribelyuk, "The number of 'magic' squares, cubes, and hypercubes", *Amer. Math. Monthly* **110**:8 (2003), 707–717. MR 2004k:05009 Zbl 1043.05501

[Benjamin and Yasuda 1999] A. T. Benjamin and K. Yasuda, "Magic 'squares' indeed!", *Amer. Math. Monthly* **106**:2 (1999), 152–156. MR 1671865 Zbl 0979.05024

[Biggs 1979] N. L. Biggs, "The roots of combinatorics", *Historia Math.* **6**:2 (1979), 109–136. MR 80h:05003 Zbl 0407.01002

[Bolt et al. 1991] B. Bolt, R. Eggleton, and J. Gilks, "The magic hexagram", *Math. Gaz.* **75**:472 (1991), 140–142.

[Calder 1949] I. R. F. Calder, "A note on magic squares in the philosophy of Agrippa of Nettesheim", *J. Warburg Courtauld Inst.* **12** (1949), 196–199.

[Cammann 1960] S. Cammann, "The evolution of magic squares in China", *J. Amer. Orient. Soc.* **80**:2 (1960), 116–124. Zbl 0102.24401

[Cammann 1961] S. Cammann, "The magic square of three in old Chinese philosophy and religion", *Hist. Relig.* **1**:1 (1961), 37–80. Zbl 0102.24402

[Cammann 1969a] S. Cammann, "Islamic and Indian magic squares, I", *Hist. Relig.* **8**:3 (1969), 181–209.

[Cammann 1969b] S. Cammann, "Islamic and Indian magic squares, II", *Hist. Relig.* **8**:4 (1969), 271–299.

[Chernick 1938] J. Chernick, "Solution of the general magic square", *Amer. Math. Monthly* **45**:3 (1938), 172–175. MR 1524224 Zbl 0018.20302

[Datta and Singh 1992] B. Datta and A. N. Singh, "Magic squares in India", *Indian J. Hist. Sci.* **27**:1 (1992), 51–120. MR 93b:01013 Zbl 0771.01002

[Gardner 2000] M. Gardner, "Some new results on magic hexagrams", *College Math. J.* **31**:4 (2000), 274–280. MR 1786804 Zbl 0995.05512

[Xin 2008] G. Xin, "Constructing all magic squares of order three", *Discrete Math.* **308**:15 (2008), 3393–3398. MR 2009e:05039 Zbl 1145.05012

bienz2@illinois.edu   *Department of Mathematics and Statistics, Elon University, Elon, NC 27244, United States*

kyokley@elon.edu   *Department of Mathematics and Statistics, Elon University, Elon, NC 27244, United States*

ccoles@elon.edu   *Department of Mathematics and Statistics, Elon University, Elon, NC 27244, United States*

msp

Trading cookies in a gambler's ruin scenario

Kuejai Jungjaturapit, Timothy Pluta, Reza Rastegar,
Alexander Roitershtein, Matthew Temba, Chad N. Vidden and Brian Wu

(Communicated by Anant Godbole)

We consider several variations of a two-person game between a "buyer" and a "seller", whose major component is a random walk of the buyer on an interval of integers. We assume a gambler's ruin scenario, where in contrast to the classical version the walker (buyer) has the option of consuming "cookies", which, when used, increase the probability of moving in the desired direction for the next step. The cookies are supplied to the buyer by the second player (seller). We determine the equilibrium price policy for the seller and the equilibrium "cookie store" location. An initial motivation for this question is provided by the popular model of "cookie" or "excited" random walks.

1. Introduction

Consider the following modification of the classical one-dimensional gambler's ruin problem [Durrett 1996; El-Shehawey 2009], where the walker has the option of consuming a *cookie* which, when used, changes transition probabilities for the next step in a desired way. The cookies are supplied to the walker (called *buyer* in what follows) by a *seller*. The buyer starts at point $a \in \mathbb{N}$ located between 0 and $b \in \mathbb{N}$, $b \geq 2$, and performs a nearest-neighbor random walk on the integer lattice $\mathbb{Z}$. If the buyer gets to point b before 0 she is rewarded with $r > 0$ dollars while if she gets to 0 before b she wins 0 dollars. Meanwhile, the seller sets up a shop somewhere on integer sites within the interval $(0, b)$. The seller sells a certain amount of cookies at a fixed price, and each cookie gives the buyer an instant probability boost in the direction of b. The walker thus always moves one step to the right with a fixed probability $p \in (0, 1)$ from regular sites and with a larger probability $p + \varepsilon \in [p, 1]$ from the store locations, if she consumes a cookie there.

The buyer seeks to maximize her expected utility function, and she can either accept the help of the *cookie service* for the offered price or reject it. Informally speaking, the goal of this paper is to determine the equilibrium price for a cookie as well as the optimal (from the perspective of the seller) placement for the store.

From the probability theory point of view, the problems that we investigate can be described collectively as an attempt to measure the gain of the walker from exploiting a reinforcing mechanism represented by *cookies*; see for instance (13) below. It is natural to study this type of problem within a game-theoretic framework, where exact features of the reinforcing mechanism are determined through the interaction between the walker and a supplier. This is in contrast to the usual *excited* or *cookie random walk* [Antal and Redner 2005; Zerner 2005] (see [Menshikov et al. 2012] for an up-to-date review and references), where the walker, as a price-taker in a large market, has no effect on determining the parameters of the cookie environment.

More specifically, we will study subgame perfect equilibria for several variants of a two-person Stackelberg game [Gibbons 1992], that is, a game where the seller takes an action first while the buyer observes the move of the seller and then acts. An action of the leader (seller) consists of setting the price for a cookie and choosing the store location, and a strategy of the follower (buyer) consists of specifying a set of seller's actions in response to which she would be willing to consume a cookie upon each visit to the store. The variations of the game that we study in the paper differ by the form of the payoff function that is assigned to the buyer. For instance, in the basic form considered in Section 3, the buyer seeks to maximize her expected earnings, while in Section 5 the buyer is risk-averse and thus also takes the extent of the risk involved in her decisions into consideration. Throughout this paper the buyer makes a simple a-priori commitment to either purchase a cookie each time when the opportunity is present or to "ignore" the store permanently, rather than devises a strategy contingent on the realization of the random walk path. It can be shown that this assumption is actually not restrictive for a risk-neutral buyer in the basic game considered in Sections 2 and 3 while, say, a risk-averse buyer considered in Section 5 might benefit from employing a policy conditional on the number of cookies currently available at the store. Intuitively, the attractiveness of the investment in cookies decreases for a risk-averse buyer as the amount of available cookies is decreasing and hence the risk involved in the investment is increasing in the course of the game. We remark that the optimization problem which the seller faces is somewhat similar to that of a monopoly whose market is a spatially nonhomogeneous Hotelling beach [Anderson et al. 1992; Hotelling 1929] with demand curve varying randomly across the population.

The game can serve as a simplified model to explore the relationship between economic agents in a risky environment, for instance a firm in an innovative and

competitive segment of a hi-tech industry and an experienced consulting company. The firm (buyer) seeks to reduce uncertainty and increase the expected profit by investing in the consulting service at a *bottleneck* point of its production line, while the consultant (seller) wants to optimize the configuration and the price of its service package.

We next define the underlying (buyer's) random walk. Fix any $p \in (0, 1)$ and let $q = 1 - p$. Fix the store placement $n \in \mathbb{N}$ and the cookie strength $\varepsilon \in [0, q]$. Let X_k and m_k denote the location of the walker and the number of cookies available at the store at time $k \in \mathbb{N} \cup \{0\}$, respectively. Formally, the pairs $(X_k, m_k)_{k \geq 0}$ form a Markov chain on $\mathbb{Z} \times (\mathbb{N} \cup \{0, \infty\})$ with transition kernel given by

$$\mathbb{P}_n(X_{k+1} = j, m_{k+1} = m \mid X_k = i, m_k = l)$$
$$= \begin{cases} p + \mathbf{1}_{\{i=n, l>0\}} \cdot \varepsilon & \text{if } j = i + 1, m = l - 1, \\ q - \mathbf{1}_{\{i=n, l>0\}} \cdot \varepsilon & \text{if } j = i - 1, m = l - 1, \\ 0 & \text{otherwise.} \end{cases}$$

Here we use the standard convention that $\infty - 1 = \infty$ and denote by $\mathbf{1}_A$ the indicator function of the event A. That is, $\mathbf{1}_A$ is either 1 or 0 according to whether A occurred or not.

The parameters m_0, p, and ε (as well as the parameters a, b, and r introduced later in this section) are considered as given exogenous variables. Let $\mathbb{P}_{a,n}$ denote the probability measure on the path of the random walk associated with the buyer starting with probability one at $X_0 = a$, while the cookie store is placed at n. Let $\mathbb{E}_{a,n}$ be the expectation operator associated with the probability measure $\mathbb{P}_{a,n}$. We will denote by P_a and E_a, respectively, the distribution and the expectation associated with the corresponding usual random walk, that is, the one with $\varepsilon = 0$.

Choose any $b \in \mathbb{N}$, $b > n$, and let

$$\mathcal{T} = \min\{T_0, T_b\} \quad \text{with } T_j = \inf\{k : X_k = j\}, \ j \in \mathbb{Z}. \tag{1}$$

Assume that $X_0 \in (0, b)$ with probability one and that 0 and b are absorbing points for the buyer's random walk; that is, $\mathbb{P}(X_{\mathcal{T}+k} = X_{\mathcal{T}} \text{ for } k \geq 0) = 1$. If the buyer visits b before 0 she is rewarded with $r > 0$ dollars, otherwise she receives 0 dollars. The strategies of the seller are represented by the pairs (c, n), where c denotes the price for a cookie which remains fixed during the game (cf. Remark 3.1). The strategies of the buyer are represented by the mappings of the pairs (c, n) into the set $\{\mathbb{P}_{a,n}, P_a\}$, where $\mathbb{P}_{a,n}$ means the decision to use the cookies whereas P_a means the decision to ignore the cookie store and proceed as a usual random walk.

The usual cookie random walk model allows cookies to be located at each site of the integer lattice. Our assumption that all the cookies are placed in the same location makes the buyer's random walk into a nearly Markovian process, and thus ensures

a more easily treatable model. In particular, the exit probabilities $\mathbb{P}_{a,n}(T_b < T_0)$ can be explicitly computed. Random walks defined by, in a sense, small, local perturbations of P_a, have been considered by many authors. In the context of excited random walks see for instance [Davis 1999; Raimond and Schapira 2010]. It turns out that, even though our underlying random walk does not exhibit as interesting a deviation from the corresponding regular random walk as the excited random walks do (compare for instance Theorem 3.5 and Remark 7.4(a)), the perturbation by a single cookie store produces many interesting quantitative effects, and its influence is not negligible even when b is taken to infinity. For instance, according to Theorem 3.5 either a supply of cookies m_0 or a reward r of the same order as b allows the seller to maintain expected revenue when b goes to infinity, ceteris paribus. The structure of the equilibrium cost is quite curious, and is discussed in detail in Remark 3.1.

The rest of the paper is organized as follows. In Sections 2 and 3 we study the basic version of the game which is described above. In Section 4 we consider a walker with initial position uniformly distributed over the interval $[1, b - 1]$. To further explore which factors are dominant in designing the equilibrium strategies of the players, we then consider buyers with utility functions different from the expected value of their earnings. In Section 5 we consider a risk-averse buyer whereas in Section 6 we study the game where the buyer is concerned not only with the expected reward but also with the expected time it takes to achieve the reward. For comparison, we then consider in Section 7 a variant of the game with the 1-excited random walk, that is, when exactly one cookie is placed everywhere on $\mathbb{Z}$. Finally, some concluding remarks are given in Section 8.

2. Basic game: Preliminaries

In this section and the next section, we consider the following scenario. Fix any integer $b \geq 2$. There is one buyer, starting the random walk at a fixed integer point $X_0 = a$ between 0 and b. There is one seller, who is seeking to maximize her expected revenue by choosing the store's location n and the price of a cookie c. There is no production cost for the seller. The seller has m cookies to sell to the buyer, either $m \in \mathbb{N}$ or $m = \infty$. The seller charges the same price for each cookie. The walker has an option to ignore (not to buy) cookies if the price is not attractive. If the buyer chooses to use the cookie she moves on $\mathbb{Z}$ according to $\mathbb{P}_{a,n}$, otherwise her motion is according to P_a. The walker seeks to maximize her expected earnings.

Thus possible actions of the seller are represented by the collection of feasible pairs (c, n), while possible strategies of the buyer are represented by the set of functions

$$B(c, n) : [0, \infty) \times \{1, \ldots, b - 1\} \to \{\mathbb{P}_{a,n}, P_a\}.$$

The walker chooses to consume or not to consume the cookies which are supplied by the seller at site n for the marginal price c, according to whether $B(c, n) = \mathbb{P}_{a,n}$ or not. We next give a formal definition of the game. Let $\Omega_b := \{2, \ldots, b-1\}$ be the set of feasible store's locations. Recall $\mathcal{T}$ from (1) and define

$$\eta_n = \sum_{i=0}^{\mathcal{T}} \mathbf{1}_{\{X_i = n\}} \quad \text{and} \quad \eta_{n,m} = \min\{\eta_n, m\}. \tag{2}$$

That is, $\eta_{n,m}$ is the total number of "successful visits" to the store (i.e., visits when the cookies are still available) by the random walk before the absorption at either 0 or b.

Definition 2.1 (game $\Gamma_{m,a}$). • $\Gamma_{m,a}$ is a two-person Stackelberg game (the first player takes an action, the second player observes the action and then moves). The (random) payoffs of the players depend on the realization of the underlying random walk (action of Nature). In order to determine their strategies, the players consider the corresponding expected payoffs.

• The strategy set of the first player (seller) is $\mathcal{S} := [0, \infty) \times \Omega_b$. Each pair $(c, n) \in \mathcal{S}$ specifies the cookie's price $c > 0$ and the store's location $n \in \Omega_b$ chosen by the seller.

• The seller moves first and communicates her action to the second player (buyer). Then the second player determines her strategy, and starts the random walk.

• Nature determines realization of the random walk.

• The strategies of the buyer are functions $B : [0, \infty) \times \Omega_b \to \{\mathbb{P}_{a,n}, P_a\}$. The buyer will either consume a cookie priced $c \in [0, \infty)$ upon each visit to a store located at $n \in \Omega_b$ or will refrain from ever making a purchase, according to whether $B(c, n) = \mathbb{P}_{a,n}$ or $B(c, n) = P_a$, respectively.

• For given cookie price $c > 0$, store location $n \in (0, b)$, response strategy B of the buyer, and realization of buyer's random walk, player's payoffs are defined as follows:

$$u_{c,n,B} := r \cdot \mathbf{1}_{\{T_b < T_0\}} - c \cdot \eta_{n,m} \cdot \mathbf{1}_{\{B(c,n) = \mathbb{P}_{a,n}\}} \quad \text{(buyer)},$$

$$v_{c,n,B} := c \cdot \eta_{n,m} \cdot \mathbf{1}_{\{B(c,n) = \mathbb{P}_{a,n}\}} \quad \text{(seller)}.$$

Notice that the payoffs are random and depend on the realization of the underlying random walk. We next specify the game solution concept invoking expected utilities which is used throughout the paper. Denote by $\mathcal{B}$ the collection of all functions from $\mathcal{S}$ to $\{\mathbb{P}_{a,n}, P_a\}$. For any pair of strategies $S = (c, n) \in \mathcal{S}$ and $B \in \mathcal{B}$ denote by $U_{S,B}$ and $V_{S,B}$, respectively, the expected payoffs of the buyer and the seller

who play according to the strategy profile (S, B). That is,

$$U_{S,B} = \begin{cases} \mathbb{E}_{a,n}(u_{c,n,B}) & \text{if } B(c,n) = \mathbb{P}_{a,n}, \\ E_a(u_{c,n,B}) & \text{if } B(c,n) = P_a, \end{cases}$$

$$V_{S,B} = \begin{cases} \mathbb{E}_{a,n}(v_{c,n,B}) = c \cdot \mathbb{E}_{a,n}(\eta_{n,m}) & \text{if } B(c,n) = \mathbb{P}_{a,n}, \\ E_a(v_{c,n,B}) = 0 & \text{if } B(c,n) = P_a. \end{cases}$$

In the next sections we will consider several variants of the above game with different payoff functions for the seller. For the basic game $\Gamma_{m,a}$ we have

$$U_{S,B} = \begin{cases} r \cdot \mathbb{P}_{a,n}(T_b < T_0) - c \cdot \mathbb{E}_{a,n}(\eta_{n,m}) & \text{if } B(c,n) = \mathbb{P}_{a,n}, \\ r \cdot P_a(T_b < T_0) & \text{if } B(c,n) = P_a. \end{cases}$$

Definition 2.2 [Gibbons 1992]. A subgame perfect equilibrium of $\Gamma_{m,a}$ is defined as a profile of strategies $(S^*, B^*) \in \mathscr{S} \times \mathscr{B}$ such that

$$U_{S^*,B^*} \geq U_{S,B^*} \quad \text{for all } S \in \mathscr{S}, \tag{3}$$

$$V_{S,B^*} \geq V_{S,B} \quad \text{for all } S \in \mathscr{S}, B \in \mathscr{B}. \tag{4}$$

More generally, (3) and (4) define a subgame perfect equilibrium for any Stackelberg two-person game with arbitrary payoffs (U, V) and strategy sets $(\mathscr{S}, \mathscr{B})$. Throughout the paper we use "equilibrium" as synonymous to the "subgame perfect equilibrium". The following remark is in order.

Remark 2.3. The assumption that neither the seller can change the price during the course of the game, nor can the buyer reconsider her decision upon an arrival to the store, might seem to be restrictive. However, it turns out that in fact this assumption does not put a real constraint on the strategies of the players. This is discussed in Remark 3.1 below, and is due to the fact that the equilibrium price for a cookie is actually independent of m, as long as $m > 0$.

For given cookie price $c > 0$ and store location $n \in (0, b)$ let $U_c(a, n)$ denote the expected payoff of the buyer who uses the cookies. That is,

$$U_c(a, n) := \mathbb{E}_{a,n}(u_{c,n,\mathbb{P}_{a,n}}) = r \cdot \mathbb{P}_{a,n}(T_b < T_0) - c \cdot \mathbb{E}_{a,n}(\eta_{n,m}).$$

The corresponding expected revenue of the seller is denoted by $V_c(a, n)$. That is,

$$V_c(a, n) := \mathbb{E}_{a,n}(v_{c,n,\mathbb{P}_{a,n}}) = c \cdot \mathbb{E}_{a,n}(\eta_{n,m}). \tag{5}$$

Thus, for fixed a and n, the seller will set the maximal possible price for each cookie. The maximal price $c^*(a, n)$ that the buyer would be willing to pay for a cookie is determined from the equation

$$U_{c^*(a,n)}(a, n) = r \cdot P_a(T_b < T_0), \tag{6}$$

where the right-hand side is the expected payoff of the buyer without cookie reinforcement. It will turn out that this equation has a unique solution for any feasible pair (a, n). The optimal location of the store $n^* = n^*(a)$ is then given as the solution of the optimization problem

$$V_{c^*(a,n^*)}(a, n^*) = \max_{n \in \Omega_b} V_{c^*(a,n)}(a, n). \tag{7}$$

We will show below (see Lemma 3.2) that $n^*(a) = a$ is the unique solution to (7). The price is determined from (6), which can be alternatively written as

$$c^*(a, n) = \frac{r \cdot \mathbb{P}_{a,n}(T_b < T_0) - r \cdot P_a(T_b < T_0)}{\mathbb{E}_{a,n}(\eta_{n,m})}. \tag{8}$$

The core result of this section is the following observation.

Theorem 2.4. *For a fixed store location n, the maximal price $c^*(a, n)$ that the buyer would be willing to pay for a cookie in a game $\Gamma_{m,a}$ is independent of the value of a.*

Proof. Given a store location $n \in (0, b)$, the maximal price $c^*(a, n)$ is determined from (6). If $n \geq a$, we have $U_{c^*(a,n)}(a, n) = P_a(T_n < T_0) \cdot U_{c^*(a,n)}(n, n)$. Thus, using the strong Markov property, identity (6) yields

$$P_a(T_n < T_0) \cdot U_{c^*(a,n)}(n, n) = r P_a(T_b < T_0) = r P_a(T_n < T_0) P_n(T_b < T_0),$$

which implies

$$U_{c^*(a,n)}(n, n) = r P_n(T_b < T_0). \tag{9}$$

If $n \leq a$, we have $U_{c^*(a,n)}(a, n) = P_a(T_n < T_b) \cdot U_{c^*(a,n)}(n, n) + r P_a(T_b < T_n)$. Hence, using again the strong Markov property, identity (6) yields

$$P_a(T_n < T_b) \cdot U_{c^*(a,n)}(n, n) + r P_a(T_b < T_n)$$
$$= r P_a(T_b < T_0) = r P_a(T_n < T_b) P_n(T_b < T_0) + r P_a(T_b < T_n),$$

which also leads to (9) in the case $n \leq a$. This completes the proof of the theorem, since (9) for $c^*(a, n)$ is independent of the value of a.   □

Proof. Let $\rho = q/p$ and recall T_n from (1). For any integer $n \in [0, a]$ we have [Durrett 1996, p. 274]:

$$P_a(T_n < T_b) = \begin{cases} \dfrac{\rho^b - \rho^a}{\rho^b - \rho^n} & \text{if } p \neq q, \\[2ex] \dfrac{b - a}{b - n} & \text{if } p = q. \end{cases} \qquad\qquad □$$

We conclude this section with the computation of $\mathbb{E}_{a,n}(\eta_{n,m})$. Let J_n (respectively, K_n) denote the probability of returning (not returning) to n after consuming a cookie at n:

$$K_n = 1 - J_n = R_n + L_n, \tag{10}$$

where

$$R_n := (p + \varepsilon) P_{n+1}(T_b < T_n),$$

$$L_n := (q - \varepsilon) P_{n-1}(T_0 < T_n).$$

We have

$$\mathbb{E}_{a,n} = \mathbb{P}_{a,n}(T_n < \mathcal{T}) \cdot \mathbb{E}_{n,n},$$

$$\mathbb{E}_{n,n} = (1 - J_n) \sum_{i=1}^{m-1} i J_n^{i-1} + m J_n^{m-1} = \frac{1 - J_n^m}{1 - J_n}. \tag{11}$$

Throughout the paper, we use the convention that if $m = \infty$ then $J^m = m J^m = 0$ for any constant $J \in (0, 1)$ in our calculations.

3. Basic game: Main results

Our main results in this section are collected in Theorem 3.3 which includes explicit results for the values of the equilibrium price and store location in $\Gamma_{m,a}$. Theorem 3.4 extends the results to the infinite interval $(-\infty, 0]$ when $\rho < 1$.

In Theorem 3.5, for the case $p = q$, we find a natural scaling of the parameters r and m when b goes to infinity. In particular, this theorem shows that an increase in cookie supply proportional to the change in the value of b allows the seller to maintain her revenue. In other words, the effect of a single store with an adequate cookie supply on the underlying random walk cannot be neglected, even asymptotically.

Finally, Theorem 3.6 establishes monotonicity of the seller's equilibrium revenue as a function of the parameter ε. The latter result is interesting because, even though the higher quality (i.e., higher strength of the cookie, ε) means higher price for a cookie, it also means that the buyer is expected to finish the game sooner and hence implies the drop in the expected amount of cookies sold.

We will frequently make use of the "decomposition according to the first step" arguments for the underlying Markov chain $(X_k, m_k)_{k \geq 0}$, in particular exploiting the following equality:

$$P_k(\mathcal{T} = T_x) = p P_{k+1}(\mathcal{T} = T_x) + q P_{k-1}(\mathcal{T} = T_x), \tag{12}$$

with $x \in \{0, b\}$ and $n = 1, \ldots, b - 1$. The recurrence relationship (12) can be equivalently stated as the martingale-type identity $E(Z_{k+1} \mid X_k) = Z_k$ for $Z_k = P_{X_k}(\mathcal{T} = T_x)$. We will denote $c^*(a, n^*(a))$ (which will turn out to be $c^*(a, a)$;

see Lemma 3.2 below) by $c^*(a)$ and refer to this value as the equilibrium price of the cookie in $\Gamma_{m,a}$. Thus, according to (8),

$$c^*(a) = \frac{r \cdot \mathbb{P}_{a,n^*(a)}(T_b < T_0) - r \cdot P_a(T_b < T_0)}{\mathbb{E}_{a,n^*(a)}(\eta_{n^*(a),m})}. \tag{13}$$

We next compute the equilibrium price $c^*(a)$. We will first calculate $U_c(n, n)$ for general $c > 0$ and $n \in (0, b)$. To simplify notation we will abbreviate $U_c(n, n)$ to $U_c(n)$ and $V_c(n, n)$ to $V_c(n)$. Recall (10). We have

$$U_c(n) = \sum_{k=1}^{m-1} \left[(r - kc) R_n J_n^{k-1} - kc L_n J_n^{k-1} \right]$$
$$+ J_n^{m-1} \left[r[(p + \varepsilon) P_{n+1}(T_b < T_0) + (q - \varepsilon) P_{n-1}(T_b < T_0)] - mc \right]. \tag{14}$$

It follows from (12) that

$$U_c(n) = \frac{R_n r (1 - J_n^{m-1})}{K_n} - c \left[\frac{(m-1) J_n^m - m J_n^{m-1} + 1}{K_n} + m J_n^{m-1} \right]$$
$$+ J_n^{m-1} r \left[P_n(T_b < T_0) + \varepsilon \left(P_{n+1}(T_b < T_0) - P_{n-1}(T_b < T_0) \right) \right].$$

Therefore, using (6) and the following identity (recall that $K_n = 1 - J_n$):

$$\frac{(m-1) J_n^m - m J_n^{m-1} + 1}{K_n} + m J_n^{m-1} = \frac{1 - J_n^m}{K_n},$$

we obtain

$$\frac{c^*(n)(1 - J_n^m)}{K_n} = \frac{R_n r (1 - J_n^{m-1})}{K_n} + J_n^{m-1} r \left[P_n(T_b < T_0) + \varepsilon(P_{n+1}(T_b < T_0) \right.$$
$$\left. - P_{n-1}(T_b < T_0)) \right] - r P_n(T_b < T_0).$$

Thus $c^*(n)$ can be expressed as

$$c^*(n) = \frac{c_1(n) + c_2(n)}{1 - J_n^m},$$

where

$$c_1(n) = r[R_n - K_n \cdot P_n(T_b < T_0)],$$
$$c_2(n) = r J_n^{m-1} \left[K_n[P_n(T_b < T_0) + \varepsilon(P_{n+1}(T_b < T_0) - P_{n-1}(T_b < T_0))] - R_n \right]$$
$$= J_n^{m-1} \left[r\varepsilon K_n[P_{n+1}(T_b < T_0) - P_{n-1}(T_b < T_0)] - c_1(n) \right].$$

We have

$$
\begin{aligned}
c_1(n) = {} & r(p+\varepsilon)P_{n+1}(T_b < T_n) \\
& - r[(p+\varepsilon)P_{n+1}(T_b < T_n) + (q-\varepsilon)P_{n-1}(T_0 < T_n)] \cdot P_n(T_b < T_0) \\
= {} & r(p+\varepsilon)P_{n+1}(T_b < T_n) \cdot P_n(T_0 < T_b) - r(q-\varepsilon)P_{n-1}(T_0 < T_n) \cdot P_n(T_b < T_0) \\
= {} & r[pP_{n+1}(T_b < T_n) \cdot P_n(T_0 < T_b) - qP_{n-1}(T_0 < T_n) \cdot P_n(T_b < T_0)] \\
& + \varepsilon r[P_{n+1}(T_b < T_n) \cdot P_n(T_0 < T_b) + P_{n-1}(T_0 < T_n) \cdot P_n(T_b < T_0)] \\
:= {} & c_{1,1}(n) + c_{1,2}(n),
\end{aligned}
$$

where the last equality serves as the definition of $c_{1,1}(n)$ and $c_{1,2}(n)$. Using the Markov property and (12), we obtain

$$
\begin{aligned}
c_{1,1}(n) = {} & r[p(1 - P_{n+1}(T_n < T_b)) \cdot P_n(T_0 < T_b) - q(1 - P_{n-1}(T_n < T_0)) \cdot P_n(T_b < T_0)] \\
= {} & r[p(P_n(T_0 < T_b) - P_{n+1}(T_0 < T_b)) - q(P_n(T_b < T_0) - P_{n-1}(T_b < T_0))] \\
= {} & r[p(P_n(T_0 < T_b) - P_{n+1}(T_0 < T_b)) - q(P_{n-1}(T_0 < T_b) - P_n(T_0 < T_b))] \\
= {} & 0
\end{aligned}
$$

and

$$
\begin{aligned}
c_{1,2}(n) = {} & \varepsilon r[(1 - P_{n+1}(T_n < T_b)) \cdot P_n(T_0 < T_b) + (1 - P_{n-1}(T_n < T_0)) \cdot P_n(T_b < T_0)] \\
= {} & \varepsilon r[1 - P_{n+1}(T_n < T_b)P_n(T_0 < T_b) - P_{n-1}(T_b < T_0)] \\
= {} & \varepsilon r[P_{n-1}(T_0 < T_b) - P_{n+1}(T_0 < T_b)] = \varepsilon r[P_{n+1}(T_b < T_0) - P_{n-1}(T_b < T_0)].
\end{aligned}
$$

Further,

$$
\begin{aligned}
c_2(n) & = J_n^{m-1}\big[r\varepsilon K_n[P_{n+1}(T_b < T_0) - P_{n-1}(T_b < T_0)] - c_1(n)\big] \\
& = -J_n^{m-1}r\varepsilon(1 - K_n) \cdot [P_{n+1}(T_b < T_0) - P_{n-1}(T_b < T_0)] \\
& = -J_n^m r\varepsilon \cdot [P_{n+1}(T_b < T_0) - P_{n-1}(T_b < T_0)].
\end{aligned}
$$

Thus

$$
c^*(n) = \frac{c_1(n) + c_2(n)}{1 - J_n^m} = r\varepsilon[P_{n+1}(T_b < T_0) - P_{n-1}(T_b < T_0)], \qquad (15)
$$

which yields

$$
c^*(n) = \begin{cases} \dfrac{r\varepsilon\rho^n(\rho^{-1} - \rho)}{1 - \rho^b} & \text{if } p \neq q, \\ 2\varepsilon r/b & \text{if } p = q. \end{cases} \qquad (16)
$$

Remark 3.1. Remarkably, $c^*(n)$ is independent of m. Furthermore, (15) implies that, given the store location n, the equilibrium price $c^*(n)$ is the unique positive constant c which makes $M_k = r \cdot P_{X_k}(T_b < T_0) - c \cdot \sum_{i=0}^{\min\{k,m\}} \mathbf{1}_{\{X_i=n\}}$ into a martingale under $\mathbb{P}_{a,n}$ with respect to the natural filtration $\mathscr{F}_k = \sigma((X_i, \mathbf{y}_i) : i \leq k)$

of the Markov chain formed by the pairs $(X_k, \mathbf{y}_k)$. Notice that $P_{X_k}(T_b < T_0)$, $k \geq 0$, is a martingale with respect to its natural filtration under P_a, but not under $\mathbb{P}_{a,n}$.

The independence of $c^*(n)$ of m is an implication of the Markov property and our assumption that the buyer is risk-neutral, and thus is concerned only with the expected value of her earnings. Using the Markov property, (8) can be rewritten as

$$c^*(n) = \frac{r \cdot \mathbb{P}_{n,n}(T_b < T_0) - r \cdot P_n(T_b < T_0)}{\mathbb{E}_{n,n}(\eta_{n,m})}.$$

The difference $\mathbb{P}_{n,n}(T_b < T_0) - r \cdot P_n(T_b < T_0)$ can be decomposed into the sum of the expected gain from using 1 cookie until either returning to the store or finishing the game. Notice that, between two successive visits to the store, the buyer's motion is described by the measure P_a. Given the possibility to reconsider her decision to use cookies upon the next return to the store, the buyer would evaluate her expected earnings again according to (6). Therefore, using the Markov property, the buyer's gain from using one cookie is, up to the multiplicative factor r,

$$(p+\varepsilon)P_{n+1}(T_b < T_n) + [(p+\varepsilon)P_{n+1}(T_n < T_b) + (q-\varepsilon)P_{n-1}(T_n < T_0)]P_n(T_b < T_0)$$
$$- pP_{n+1}(T_b < T_n) + [pP_{n+1}(T_n < T_b) + qP_{n-1}(T_n < T_0)]P_n(T_b < T_0)$$
$$= \varepsilon P_{n+1}(T_b < T_n) + \varepsilon[P_{n+1}(T_n < T_b) - P_{n-1}(T_n < T_0)]P_n(T_b < T_0)$$
$$= \varepsilon[P_{n+1}(T_b < T_0) - \varepsilon P_{n-1}(T_b < T_0)],$$

in agreement with (15).

As we already mentioned in Remark 2.3, the fact that $c^*(n)$ is independent of m implies that the buyer would not change her decision regarding the use of cookies during the course of the game. This implies that the equilibrium price policy for the seller is to maintain a fixed cookie price throughout the game even if the buyer were allowed to change it according to the number of the cookies left in stock. The fact that the price $c^*(a)$ is a multiple of the boost ε is not surprising, though it is not trivial a priori and is interesting.

We are now in a position to find the seller's expected revenue with the store located at n. For an arbitrary $c > 0$, write, using (11),

$$V_c(n) = c \cdot \mathbb{E}_{n,n}(\eta_{n,m}) = \frac{c(1 - J_n^m)}{1 - J_n}. \tag{17}$$

Recall the convention $J_n^m = m J_n^m = 0$ for $m = \infty$. We have:

Lemma 3.2. *For a fixed starting point of the buyer a, the unique subgame perfect location of the cookie store is at $n^*(a) = a$.*

Proof. The strong Markov property and Theorem 2.4 imply that

$$V_{c^*(a,n)}(a, n) = P_a(T_n < \mathcal{T}) \cdot V_{c^*(n)}(n), \tag{18}$$

where $\mathcal{T}$ is defined in (1). For real $x \in (0, b)$ define

$$
J(x) = \begin{cases} (p + \varepsilon)\left(1 - \dfrac{1}{b - x}\right) + (q - \varepsilon)\left(1 - \dfrac{1}{x}\right) & \text{if } p = q, \\[2ex] (p + \varepsilon)\dfrac{\rho^b - \rho^{x+1}}{\rho^b - \rho^x} + (q - \varepsilon)\dfrac{\rho^{x-1} - 1}{\rho^x - 1} & \text{if } p \neq q. \end{cases}
$$

For real numbers $x \in (0, b)$ define

$$
f_{\rho,m}(x) = \begin{cases} \dfrac{1}{x} \cdot \dfrac{1 - J^m(x)}{1 - J(x)} & \text{if } x \geq a \text{ and } \rho = 1, \\[2ex] \dfrac{1}{b-x} \cdot \dfrac{1 - J^m(x)}{1 - J(x)} & \text{if } x \leq a \text{ and } \rho = 1, \\[2ex] \dfrac{\rho^x}{\rho^x - 1} \cdot \dfrac{1 - J^m(x)}{1 - J(x)} & \text{if } x \geq a \text{ and } \rho \neq 1, \\[2ex] \dfrac{\rho^x}{\rho^b - \rho^x} \cdot \dfrac{1 - J^m(x)}{1 - J(x)} & \text{if } x \leq a \text{ and } \rho \neq 1. \end{cases}
$$

Then $J_n = J(n)$, where J_n is given by (10). It follows from (16), (17), and (18) that $f_{\rho,m}(x)$ differs from $V_{c^*(x)}(a, x)$ only by a positive constant multiplicative factor on both the intervals $[1, a]$ and $[a, n]$. Considering the sign of the derivative $f'_{\rho,m}(x)$ and using the fact that

$$
\left(\frac{1 - J^m(x)}{1 - J(x)}\right)' = J'(x)\sum_{k=0}^m k J^{k-1}(x),
$$

it is easy to verify that, if the lemma is true for $m = \infty$, it is true for any $m \in \mathbb{N}$. It is then routine to check, using the first derivative test, that $f_{\rho,\infty}(x)$ (and hence $V_{c^*(x)}(a, x)$) attains its maximum when $x = a$ for any $\rho > 0$. The proof of the lemma is completed.   □

Note that, in the extreme case $\varepsilon = 1 - p$, any location $n \leq a$ will have the same effect from perspective of the buyer. Thus, in that case, the seller is only concerned with optimizing the chances of the buyer to ever visit the store. We summarize our results for the subgame perfect equilibrium strategy $(c^*(a), n^*(a))$ of the seller and her corresponding revenue in the following statement.

Theorem 3.3. *Consider a game $\Gamma_{m,a}$.*

(a) *For a fixed starting point of the buyer a, the unique subgame perfect equilibrium location of the store is at $n^* = a$.*

(b) *The subgame perfect equilibrium cost $c^*(a)$ is given by (16) with $n = a$; thus*

$$
c^*(a) = \begin{cases} \dfrac{r\varepsilon\rho^a(\rho^{-1} - \rho)}{1 - \rho^b} & \text{if } p \neq q, \\[2ex] 2\varepsilon r/b & \text{if } p = q. \end{cases}
$$
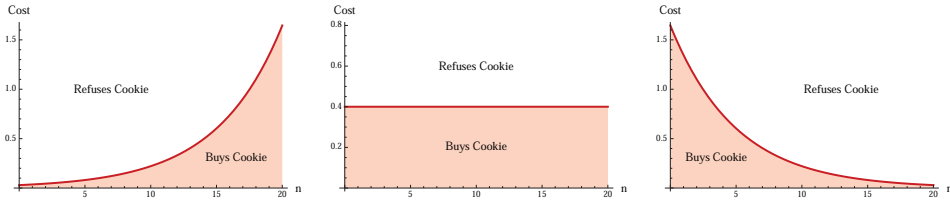
Figure 1. Sketch of the graph of $c^*(n)$ and buyer's equilibrium policy. From left to right, the graphs exemplify the cases $\rho > 1$, $\rho = 1$, $\rho < 1$. In all cases, $b = r = 10$ and $\varepsilon = 0.2$.

In particular, $c^*(a)$ is independent of the value of m.

(c) *The expected revenue of the seller at equilibrium is given by*

$$V^*(a) := V_{c^*(a)}(a) = \frac{c^*(a)(1 - J_a^m)}{1 - J_a}.$$

Figure 1 provides a graphical representation of the equilibrium strategy of the buyer in the first quadrant of the plane (c, n), where each point corresponds to an available strategy of the seller. In all three cases illustrated, $b = r = 10$ and $\varepsilon = 0.2$.

Assuming $q > p$, one can consider a version of the game on the interval $(-\infty, 0]$ with a reward $r > 0$ given to the walker when (and if) she arrives at 0. The equilibrium strategies for the game on $(-\infty, 0)$ can be formally obtained from the corresponding results for a finite interval by replacing a with $a + b$ and taking the limit as $b \to \infty$. We state this as follows. Let

$$J_a^{\text{inf}} = (p + \varepsilon)P_{a+1}(T_a < T_0) + (q - \varepsilon)P_{a-1}(T_a < T_{-\infty})$$

$$= (p + \varepsilon)\frac{\rho^{a+1} - 1}{\rho^a - 1} + (q - \varepsilon)\rho^{-1} = 1 + \varepsilon(\rho - \rho^{-1}) - (p + \varepsilon)\frac{\rho - 1}{1 - \rho^a}.$$

Theorem 3.4. *Consider the variant of $\Gamma_{m,a}$ where the buyer's random walk is taking place on the infinite interval $(-\infty, 0]$, the site 0 is the unique absorbing point for the random walk, $\rho > 1$, the buyer is rewarded with $r > 0$ dollars when (and if) she reaches 0, and the buyer's starting point is a fixed constant $a \in (-\infty, 0)$.*

(a) *The equilibrium location of the store is at $n_{\text{inf}}^* = a$.*

(b) *For a given price for a cookie c, provided that the buyer will use the cookies, the expected revenue of the seller is given by*

$$V_c^{\text{inf}}(a) = c \cdot \frac{1 - (J_a^{\text{inf}})^m}{1 - J_a^{\text{inf}}}.$$

(c) *The equilibrium cost $c_{\text{inf}}^*(a)$ is given by $c_{\text{inf}}^*(a) = r\varepsilon\rho^a(\rho - \rho^{-1})$. In particular, $c_{\text{inf}}^*(a)$ is independent of the value of m.*

The explicit formulas provided by Theorem 3.3 allow one also to study how the main characteristics of the buyer-seller game depend on the parameters b and ε. In the next theorem we find natural scalings for the parameters m and r when $p = q$ and b goes to infinity. The scaling factors turn out to be of order b, indicating that the effect of the cookie store on the simple random walk is considerably large.

Theorem 3.5. (a) *For any $a \in \mathbb{N}$ and $m \in \mathbb{N} \cup \{\infty\}$, if $\lim_{b\to\infty} b^{-1} r(b) = \alpha$ for some constant $\alpha \in (0, \infty)$, we have $\lim_{b\to\infty} c^*(a) = 2\varepsilon\alpha$.*

(b) *For any $x > 0$, if $r > 0$ and $\lim_{b\to\infty} b^{-1} \cdot m(b) = \beta$ for some constant $\beta \in (0, \infty)$ (and giving $\lfloor bx \rfloor$ its usual meaning, $\max\{n \in \mathbb{N} : n \leq bx\}$), we have*

$$\lim_{b\to\infty} V^*(\lfloor bx \rfloor) = 2\varepsilon r \cdot \frac{1 - e^{-\beta K_0}}{K_0}, \quad \text{where } K_0 = \frac{1}{2}\left(\frac{1 + 2\varepsilon}{1 - x} + \frac{1 - 2\varepsilon}{x}\right).$$

We next investigate the equilibrium revenue of the seller $V^*(a)$ as a function of the parameter ε. On one hand, the seller provides cookies creating a positive reinforcement to the random walk to terminate at b. On the other hand, in order to increase consumption of cookies, she is interested in keeping the walker in the game as long as possible. The following result shows that, in the trade-off between the equilibrium price $c^*(a)$ (increasing function of ε) and the expected number of visits to the store (decreasing function of ε), the former is the dominant factor for establishing the equilibrium policy of the seller.

Theorem 3.6. $V^*(a)$ *is an increasing function of the parameter ε.*

Proof. Observe that, for any $\rho > 0$, $c^*(a)$ has the form $c^*(a) = C\varepsilon$ where $C > 0$ does not depend on ε. Therefore, by Theorem 3.3,

$$\frac{\partial V^*(a)}{\partial \varepsilon} = \frac{C(1 - J_a^m)}{1 - J_a} + \frac{\partial J_a}{\partial \varepsilon} \cdot \frac{C\varepsilon(1 - m J_a^{m-1} + (m-1)J_a^m)}{(1 - J_a)^2}. \tag{19}$$

According to (10),

$$\frac{\partial J_a}{\partial \varepsilon} = P_{a+1}(T_a < T_b) - P_{a-1}(T_a < T_0) > \frac{J_a - 1}{\varepsilon}.$$

Furthermore,

$$\frac{1 - m J_a^{m-1} + (m-1)J_a^m}{(1 - J_a)^2} = \frac{\partial}{\partial J_a}\left(\frac{1 - J_a^m}{1 - J_a}\right) = \sum_{k=1}^{m} k J_a^{k-1} > 0.$$

Therefore, replacing $\frac{\partial J_a}{\partial \varepsilon}$ with $\frac{J_a - 1}{\varepsilon}$ in (19), we obtain

$$\frac{\partial V^*(a)}{\partial \varepsilon} > \frac{C(1 - J_a^m)}{1 - J_a} - \frac{C(1 - m J_a^{m-1} + (m-1)J_a^m)}{1 - J_a} = Cm J_a^{m-1} \geq 0.$$

This completes the proof of the theorem. □

4. Population of buyers. Randomized entry point for the buyer

In this section we aim to find the equilibrium policy (c, n) for a single seller dealing with a population of walkers. Notice that, according to Theorem 2.4, once the store is placed, the equilibrium price for a cookie is independent of the buyer's entry point and therefore is determined by the store placement only.

Assume that the buyers are independent of each other, and the starting position of each buyer is distributed uniformly on $\{1, \ldots, b-1\}$. Further, assume that the path of the random walk associated with the buyer with $X_0 = a$ is distributed according to $\mathbb{P}_{a,n}$ with $m = \infty$. It then follows from (2) that the problem is basically equivalent to its analogue with a single buyer whose initial position is uniformly distributed over the integers within $(0, b)$. In what follows we will therefore consider a slightly more general scenario, formally allowing $m < \infty$.

Definition 4.1. The game $\Gamma_{m,\text{unif}}$ is the same as $\Gamma_{m,a}$, except that the buyer starts her random walk at a random integer point X_0, uniformly distributed over $(0, b)$.

Let $V_c^{\text{unif}}(n)$ denote the expected revenue of a seller whose store is located at site n. For $n \in [1, b-1]$ we have

$$V_c^{\text{unif}}(n) = \frac{1}{b-1} \sum_{a=1}^{b-1} V_c(a, n) = \frac{V_c(n)}{b-1} \sum_{a=1}^{b-1} P_a(T_n < \mathcal{T})$$

$$= \frac{V_c(n)}{b-1} \left[1 + \sum_{a=1}^{n-1} P_a(T_n < T_0) + \sum_{a=n+1}^{b-1} P_a(T_n < T_b) \right],$$

with the usual convention that the last sum vanishes if $n + 1 > b - 1$. For $p = q$ we obtain

$$1 + \sum_{a=1}^{n-1} P_a(T_n < T_0) + \sum_{a=n+1}^{b-1} P_a(T_n < T_b)$$

$$= \sum_{a=1}^{n-1} \frac{a}{n} + \sum_{a=n}^{b-1} \frac{b-a}{b-n} = \frac{n-1}{2} + b - \frac{(b-1)b - (n-1)n}{2(b-n)} = \frac{n-1}{2} + b - \frac{b+n-1}{2} = \frac{b}{2}.$$

For $p \neq q$ we obtain

$$1 + \sum_{a=1}^{n-1} P_a(T_n < T_0) + \sum_{a=n+1}^{b-1} P_a(T_n < T_b) = \sum_{a=1}^{n-1} \frac{\rho^a - 1}{\rho^n - 1} + \sum_{a=n}^{b-1} \frac{\rho^b - \rho^a}{\rho^b - \rho^n}$$

$$= \frac{\frac{\rho^n - 1}{\rho - 1} - 1}{\rho^n - 1} - \frac{n-1}{\rho^n - 1} + \frac{\rho^b(b-n)}{\rho^b - \rho^n} - \frac{\frac{\rho^b - \rho^n}{\rho - 1}}{\rho^b - \rho^n}$$

$$= -\frac{n}{\rho^n - 1} + \frac{\rho^b(b-n)}{\rho^b - \rho^n}.$$

We summarize this calculation in the following lemma. Recall $V_c(n)$ from page 199.

Lemma 4.2. *Consider a game $\Gamma_{m,\text{unif}}$.*

(a) *If $p = q$, we have*

$$V_c^{\text{unif}}(n) = \frac{V_c(n)b}{2(b-1)}.$$

(b) *If $p \neq q$, we have*

$$V_c^{\text{unif}}(n) = \frac{V_c(n)}{b-1} \cdot \left(-\frac{n}{\rho^n - 1} + \frac{\rho^b(b-n)}{\rho^b - \rho^n} \right).$$

Let $V_{\text{unif}}^*(n)$ denote the maximal expected revenue in $\Gamma_{m,\text{unif}}$ of the store located at $n \in (0, b)$. That is, $V_{\text{unif}}^*(n) = V_{c^*(n)}^{\text{unif}}(n)$, where $c^*(n)$ is defined in (13). Recall J_n from (10). Combining Lemma 4.2 with (17) and Theorem 3.3(a), we obtain:

Corollary 4.3. *Consider a game $\Gamma_{m,\text{unif}}$.*

(a) *If $p = q$, we have*

$$V_{\text{unif}}^*(n) = \frac{\varepsilon r(1 - J_n^m)}{(b-1)(1 - J_n)}.$$

(b) *If $p \neq q$, we have*

$$V_{\text{unif}}^*(n) = \frac{r\varepsilon\rho^n(\rho^{-1} - \rho)}{1 - \rho^b} \cdot \frac{1 - J_n^m}{(b-1)(1 - J_n)} \cdot \left(-\frac{n}{\rho^n - 1} + \frac{\rho^b(b-n)}{\rho^b - \rho^n} \right).$$

Corollary 4.4. *Let a real number $t \in [0, 1]$ be fixed and let $(n_b)_{b \in \mathbb{N}}$ be any sequence of integers such that $\lim_{b \to \infty} n_b/b = t$. Then*

$$\lim_{b \to \infty} \frac{V_{\text{unif}}^*(n_b)}{V^*(n_b)} = \begin{cases} 1 - t & \text{if } \rho > 1, \\ 1/2 & \text{if } \rho = 1, \\ t & \text{if } \rho < 1, \end{cases}$$

where both $V_{\text{unif}}^*(n_b)$ and $V^*(n_b)$ are computed for arbitrary but always the same values of m and r, which may or may not depend on b.

We turn now to the study of the equilibrium location of the cookie store under the assumption given in Definition 4.1. For $p = q$ we have

$$J_n = (p + \varepsilon)\left(1 - \frac{1}{b-n}\right) + (q - \varepsilon)\left(1 - \frac{1}{n}\right) = 1 - \left(\frac{p + \varepsilon}{b-n} + \frac{q - \varepsilon}{n}\right).$$

For a real number $x \in (0, b)$ let

$$f(x) = \left(\frac{p + \varepsilon}{b-x} + \frac{q - \varepsilon}{x}\right).$$

Then $\lim_{x \to 0} f(x) = \lim_{x \to b} f(x) = +\infty$ and $f(x)$ is minimal over $(0, b)$ when

$f'(x) = 0$, that is, when

$$\frac{p + \varepsilon}{(b - x)^2} = \frac{q - \varepsilon}{x^2}.$$

This yields $(p - q + 2\varepsilon)x^2 + 2bx(q - \varepsilon) - b^2(q - \varepsilon) = 0$. The unique root of this equation which belongs to the interval $(0, b)$ is given by

$$x_0(\varepsilon) = \frac{-2b(q - \varepsilon) + 2b\sqrt{(q - \varepsilon)^2 + (p - q + 2\varepsilon)(q - \varepsilon)}}{2(p - q + 2\varepsilon)}$$

$$= b\frac{-(1 - 2\varepsilon) + \sqrt{1 - 4\varepsilon^2}}{4\varepsilon}. \tag{20}$$

For the equilibrium location of the store n^*_{unif} we have $|n^*_{\text{unif}} - x_0(\varepsilon)| < 1$. Notice that $\lim_{\varepsilon \to 0} x_0(\varepsilon) = b/2$, $x_0(1/2) = 0$, and

$$x_0'(\varepsilon) = \frac{1 - \sqrt{1 - 4\varepsilon^2}}{4\varepsilon^2} - \frac{1}{\sqrt{1 - 4\varepsilon^2}} = \frac{1}{1 + \sqrt{1 - 4\varepsilon^2}} - \frac{1}{\sqrt{1 - 4\varepsilon^2}} < 0.$$

We next examine the optimal location of the store for the case $p \neq q$ and $m = \infty$. Let $A = (p + \varepsilon)(1 - \rho)$ and $B = (q - \varepsilon)(1 - \rho^{-1})$. Then it is routine to check, using the first derivative test, that $|n^*_{\text{unif}} - x_0(\varepsilon)| < 1$, where $x_0(\varepsilon) \in (0, b)$ is the unique solution of the equation

$$x \ln \rho \cdot (A + B\rho^b) + [A + B\rho^b - Bb\rho^b \ln \rho] = \rho^x(A + B). \tag{21}$$

It is not hard to check that

$$\lim_{\varepsilon \to 0} x_0(\varepsilon) = \frac{b\rho^b}{\rho^b - 1} - \frac{1}{\ln \rho} > 0, \quad \lim_{\varepsilon \to q} x_0(\varepsilon) = 0, \; x_0'(\varepsilon) < 0.$$

The value of $x_0(\varepsilon)$ that solves (21) gives us insight as to which point will maximize the seller's expected revenue. We summarize the calculations above as follows.

Lemma 4.5. *Consider a game $\Gamma_{m,\text{unif}}$. If $p \neq q$, assume in addition that $m = \infty$. Then $|n^*_{\text{unif}} - x_0(\varepsilon)| < 1$ where, for $p = q$, $x_0(\varepsilon)$ is given by (20), while for $p \neq q$, $x_0(\varepsilon)$ is determined as the unique positive solution to (21).*

Corollary 4.6. *Under the conditions of Lemma 4.5, $x_0(\varepsilon)$ is a decreasing function of the parameter ε. Furthermore*:

(1) $\lim_{\varepsilon \to q} x_0(\varepsilon) = 0$;

(2) $\lim_{\varepsilon \to 0} x_0(\varepsilon) = b/2$ *for $p = q$*;

(3) *For a fixed $\rho > 0$ and for $p \neq q$ and $m = \infty$, we have*

$$\lim_{\varepsilon \to 0} x_0(\varepsilon) = \frac{b\rho^b}{\rho^b - 1} - \frac{1}{\ln \rho} > 0.$$

Corollary 4.7. *Under the conditions of Lemma 4.5, for fixed ρ, r, and $\varepsilon > 0$ we have*:

(1) *The quotient $x_0(\varepsilon)/b$ is a decreasing, constant, or increasing function of b according to whether ρ is less than, equal to, or greater than one.*

(2) *If $\rho > 1$ (and $m = \infty$), then $\lim_{b \to \infty} n^*_{\text{unif}} = b - 1$.*

(3) *If $\rho < 1$ (and $m = \infty$), then $\lim_{b \to \infty} x_0(\varepsilon) = \hat{x}_\varepsilon$ where $\hat{x}_\varepsilon$ is the unique positive solution to the equation $A(1 + x \ln \rho) = \rho^x(A + B)$.*

Corollary 4.6 implies that the range for the equilibrium store placement computed for all possible values of ε and fixed b, r, and ρ, is the whole interval $(0, n_{\max})$ for some integer $n_{\max} \in (0, b)$. This is in stark contrast with the basic model, where the buyer's initial position is the major factor influencing the seller's decision regarding the optimal store placement. This can be heuristically explained recalling that the optimal store location is determined in the trade-off between the equilibrium price for a cookie and the expected number of visits to the store. The assumption that the buyer's entry point is spread uniformly over $(0, b)$ smooths out the influence of the "accessibility" factor, and therefore implies that the price optimization gets more weight than it had for a "deterministically starting" buyer.

5. Risk aversion

In this section we aim to compare the two-person game considered in Section 2 with a version where the buyer is risk-averse when making decisions under uncertainty. The main result of the section is stated in Theorem 5.2.

In this section we consider the following variation of the basic game.

Definition 5.1. The game $\Gamma_{m,\text{ra}}$ is the same as same as $\Gamma_{m,a}$ except that the buyer's goal is to maximize her utility function given, for some fixed constants $A \geq 0$ and $\alpha \in (0, 1)$, by

$$U_c^{\text{ra}}(a, n) = \mathbb{E}_{a,n}(x - A\alpha^x), \tag{22}$$

where $x = r \cdot \mathbf{1}_{\{\mathcal{T} = T_b\}} - c \cdot \eta_{n,m}$ is the total earnings of the buyer during the game (possibly negative). Here, as before, c is the price taken by a seller for a cookie, m is the number of cookies available at the store, and $\eta_{n,m}$ is introduced in (2).

The individual utility function in the form (22) is a particularly popular choice in economics literature, used for modeling risk-averse behavior. See for instance [Bell 1988; Bell and Fishburn 2001] for its axiomatic characterization. The utility function of the seller in this section is the same as the one in Section 3, namely the expected payment of the buyer to the seller, $\mathbb{E}_{a,n}(c \cdot \eta_{n,m})$. That is, in contrast to the buyer, the seller is risk-neutral.

The equilibrium price for a cookie $c_{\text{ra}}^*(a, n)$ can be determined as a solution for unknown variable c to the equation

$$U_c^{\text{ra}}(a, n) = (r - A\alpha^r) \cdot P_a(T_b < T_0) - A \cdot P_a(T_0 < T_b), \tag{23}$$

which is the counterpart of (6) for a risk-averse buyer. Notice that, according to (22), $U_c^{\text{ra}}(a, n)$ is a decreasing function of the parameter c with $\lim_{c \to \infty} U_c^{\text{ra}}(a, n) = -\infty$. Furthermore, $U_0^{\text{ra}}(a, n) = (r - A\alpha^r) \cdot \mathbb{P}_{a,n}(T_b < T_0) - A \cdot \mathbb{P}_{a,n}(T_0 < T_b)$, and hence

$$U_0^{\text{ra}}(a, n) - (r - A\alpha^r) \cdot P_a(T_b < T_0) - A \cdot P_a(T_0 < T_b)$$
$$= [\mathbb{P}_{a,n}(T_b < T_0) - P_a(T_b < T_0)] \cdot [r + A(1 - \alpha^r)] > 0.$$

Therefore, (23) has a unique positive solution. The main result of this section is stated in the following theorem. Recall $c^*(n)$ from (13).

Theorem 5.2. *Consider a game* $\Gamma_{m,\text{ra}}$. *Then* $c_{\text{ra}}^*(a, n) \leq c^*(n)$.

Proof. Let $R_{\mathcal{J}} = r\mathbf{1}_{\{T_b = \mathcal{J}\}}$. According to (23), $c_{\text{ra}}^*(a, n)$ is the unique solution for c to the equation

$$E_a(x - A\alpha^x) = \mathbb{E}_{a,n}[(R_{\mathcal{J}} - c\eta_{n,m}) - A\alpha^{R_{\mathcal{J}} - c\eta_{n,m}}].$$

To avoid using two different expectation functionals, namely E_a and $\mathbb{E}_{a,n}$, in the same equation, we can enlarge the probability space, where the random walk $(X_n)_{n \geq 0}$ is defined, to include a random walk $Y = (Y_n)_{n \geq 0}$ which is independent of $(X_n)_{n \geq 0}$, starts at $Y_0 = a$ with probability one, ignores cookies, and is distributed according to P_a. We will assume that the second walker is also rewarded r dollars if she reaches b. Let y denote buyer's earnings; that is, $y = r \cdot \mathbf{1}_{\{Y \text{ hits } b \text{ before } 0\}}$. Using this notation we obtain the equation for $c_{\text{ra}}^*(a, n)$ in the following form:

$$\mathbb{E}_{a,n}(y - A\alpha^y) = \mathbb{E}_{a,n}[(R_{\mathcal{J}} - c_{\text{ra}}^*(a, n) \cdot \eta_{n,m}) - A\alpha^{R_{\mathcal{J}} - c_{\text{ra}}^*(a,n) \cdot \eta_{n,m}}].$$

The latter is equivalent to

$$c_{\text{ra}}^*(a, n) = \frac{\mathbb{E}_{a,n}(R_{\mathcal{J}} - y)}{\mathbb{E}_{a,n}(\eta_{n,m})} - A \cdot \frac{\mathbb{E}_{a,n}[\alpha^{R_{\mathcal{J}} - c_{\text{ra}}^*(a,n) \cdot \eta_{n,m}} - \alpha^y]}{\mathbb{E}_{a,n}(\eta_{n,m})}$$
$$= c^*(n) - A \cdot \frac{\mathbb{E}_{a,n}[\alpha^{R_{\mathcal{J}} - c_{\text{ra}}^*(a,n) \cdot \eta_{n,m}} - \alpha^y]}{\mathbb{E}_{a,n}(\eta_{n,m})}. \tag{24}$$

Therefore, the statement of the theorem is equivalent to the claim that (recall that two random walks under consideration are independent of each other)

$$\mathbb{E}_{a,n}[\alpha^{R_{\mathcal{J}} - y - c_{\text{ra}}^*(a,n) \cdot \eta_{n,m}}] > 1.$$

Hence it suffices to show that the above inequality holds. Toward this end, observe that $f(c) := \mathbb{E}_{a,n}(\alpha^{R_{\mathcal{J}} - y - c\eta_{n,m}})$ is an increasing function of the parameter c. Therefore, if it were the case that $c^*(n) \leq c_{\text{ra}}^*(a, n)$ and $\mathbb{E}_{a,n}[\alpha^{R_{\mathcal{J}} - y - c_{\text{ra}}^*(a,n) \cdot \eta_{n,m}}] \leq 1$, we

would also have

$$\mathbb{E}_{a,n}[\alpha^{R_{\mathcal{T}}-y-c^*(n)\cdot\eta_{n,m}}] \leq 1. \tag{25}$$

It follows from (8) that $\mathbb{E}_{a,n}[R_{\mathcal{T}} - y - c^*(n) \cdot \eta_{n,m}] = 0$, and hence (25) violates Jensen's inequality for the convex function α^x. The proof of the theorem is therefore completed. $\qquad\square$

The intuitive explanation for the above result is as follows. While the walker described by $(Y_n)_{n\geq 0}$ is risk-neutral and uses the expected earnings as her utility function, the first walker is "more skeptical" (risk-averse) and therefore she effectively values the expected earning less than its nominal value.

It is not hard to check that the proof of Theorem 2.4 goes through and hence its conclusion is in force for $\Gamma_{m,\mathrm{ra}}$. That is, for a fixed store location n, the maximal price $c_{\mathrm{ra}}^*(a, n)$ that the buyer would be willing to pay for a cookie is independent of the value of a. This can also be derived directly from (24). Indeed, using the fact that

$$\mathbb{E}_{a,n}\left[(\alpha^{R_{\mathcal{T}}-c_{\mathrm{ra}}^*(a,n)\cdot\eta_{n,m}} - \alpha^y)\mathbf{1}_{\{\eta_{n,m}=0\}}\right] = E_a(\alpha^{R_{\mathcal{T}}} \cdot \mathbf{1}_{\{\eta_{n,m}=0\}}) - E_a(\alpha^{R_{\mathcal{T}}} \cdot \mathbf{1}_{\{\eta_{n,m}=0\}}) = 0$$

and the Markov property, we obtain from (24) the following equation independent of a:

$$
\begin{aligned}
c_{\mathrm{ra}}^*(a, n) &= c^*(n) - A \cdot \frac{\mathbb{E}_{a,n}[\alpha^{R_{\mathcal{T}}-c_{\mathrm{ra}}^*(a,n)\cdot\eta_{n,m}} - \alpha^y]}{\mathbb{E}_{a,n}(\eta_{n,m})} \\
&= c^*(n) - A \cdot \frac{\mathbb{E}_{n,n}[\alpha^{R_{\mathcal{T}}-c_{\mathrm{ra}}^*(a,n)\cdot\eta_{n,m}} - \alpha^y]}{\mathbb{E}_{n,n}(\eta_{n,m})}.
\end{aligned}
$$

We can therefore simplify the notation $c_{\mathrm{ra}}^*(a, n)$ to $c_{\mathrm{ra}}^*(n)$. Since $\eta_{n,m}$ and $R_{\mathcal{T}}$ are independent random variables under $\mathbb{P}_{n,n}$, we obtain that $c_{\mathrm{ra}}^*(n)$ is the unique solution of the equation

$$c_{\mathrm{ra}}^*(n) = c^*(n) - A \cdot \frac{\mathbb{E}_{n,n}(\alpha^{-c_{\mathrm{ra}}^*(n)\cdot\eta_{n,m}}) \cdot \mathbb{E}_{n,n}(\alpha^{R_{\mathcal{T}}}) - E_n(\alpha^{R_{\mathcal{T}}})}{\mathbb{E}_{n,n}(\eta_{n,m})}. \tag{26}$$

Though it seems impossible to determine the optimal location of the store from this equation analytically, it can be useful for numerical analysis since all the expectations appearing in the equation can be computed explicitly. We remark that, in virtue of Theorem 5.2, (26) yields the following lower bound for $c_{\mathrm{ra}}^*(n)$:

$$c_{\mathrm{ra}}^*(n) \geq c^*(n) - A \cdot \frac{\mathbb{E}_{n,n}(\alpha^{-c^*(n)\cdot\eta_{n,m}}) \cdot \mathbb{E}_{n,n}(\alpha^{R_{\mathcal{T}}}) - E_n(\alpha^{R_{\mathcal{T}}})}{\mathbb{E}_{n,n}(\eta_{n,m})}. \tag{27}$$

The right-hand side is negative and thus the bound is trivial for A large enough. When A approaches infinity, $c_{\mathrm{ra}}^*(n)$ converges to $c_\infty(n) > 0$ which is uniquely determined from the equation $\mathbb{E}_{n,n}(\alpha^{-c_\infty(n)\cdot\eta_{n,m}}) \cdot \mathbb{E}_{n,n}(\alpha^{R_{\mathcal{T}}}) = E_n(\alpha^{R_{\mathcal{T}}})$.

6. Time is money

We next consider a model where the buyer values not only the size of the reward but also the time needed to achieve this reward. Time thus represents an opportunity cost of participating in the cookie game. For simplicity, we do not assume that the payoff is directly discounted or is subject to a "bias for the present" factorization, as say in [O'Donoghue and Rabin 1999]. More precisely, we impose in this section the following assumption regarding the buyer's utility function.

Definition 6.1. The game $\Gamma_{m,\text{time}}$ is the same as $\Gamma_{m,a}$ except that the buyer's goal is to maximize her utility function given, for a fixed constant $\Lambda > 0$, by

$$U_c^{\text{time}}(a, n) = \mathbb{E}_{a,n}(x - \Lambda \mathcal{T}),$$

where $x = r \cdot \mathbf{1}_{\{\mathcal{T}=T_b\}} - c \cdot \eta_{n,m}$ is the total earning of the buyer during the game (possibly negative).

Our main result in this section is stated in Theorem 6.2, where the equilibrium price for a cookie is determined. The equilibrium cost structure can be then in principle used for finding the optimal store location. In general, the optimal placement does not necessarily coincide with the starting point of the buyer. For instance, if Λ is large enough, the buyer might be better off avoiding the use of the cookies (at any positive price) in hopes of finishing the game quickly by exiting $[0, b]$ from the left. It can be shown that the optimal placement depends not only on the entry point a and the relationship between the reward r and the "implicit cost" Λ, but also on the number of cookies initially available at the store, m. Since there are many possible scenarios depending on the values of all the parameters involved, we will not pursue details here.

Let $c_{\text{time}}^*(n)$ be the equilibrium price for a cookie $\Gamma_{m,\text{time}}$ when the store is placed at $n \in (0, b)$. Similarly to (2), we define

$$\eta_n(k) = \sum_{i=0}^{\min\{m,k\}} \mathbf{1}_{\{X_i=n\}}. \tag{28}$$

Notice that $\eta_n(\mathcal{T}) = \eta_n(\mathcal{T} - 1) = \eta_{n,m}$. We have:

Theorem 6.2. *Consider a game $\Gamma_{m,\text{time}}$. Then*

$$c_{\text{time}}^*(n) = \begin{cases} \dfrac{r\varepsilon\rho^n(\rho^{-1} - \rho)}{1 - \rho^b} - \dfrac{\Lambda\varepsilon}{p - q} \cdot \left(\dfrac{b\rho^n(\rho^{-1} - \rho)}{1 - \rho^b} - 2 \right) & \text{if } p \neq q, \\ 2r\varepsilon/b - 2\varepsilon\Lambda(b - 2n) & \text{if } p = q. \end{cases}$$

A negative value of $c_{\text{time}}^*(n)$ indicates that the walker will refrain from using cookies regardless of the price, and hence the seller is better off not opening the store at location n.

Proof. (a) If $p \neq q$, let

$$M_k = X_k - k \cdot (p-q) - 2\varepsilon \cdot \eta_n(k-1), \quad k \geq 0,$$

with the agreement that $\eta_n(-1) = 0$. Then $(M_k)_{k \geq 0}$ is martingale with respect to the natural filtration of the Markov chain formed by the pairs $(X_k, m_k)_{k \geq 0}$, where m_k is the number of cookies left at the store by time k, as defined in Section 1. By the optional stopping theorem (see, for instance, Theorem 7.5 in [Durrett 1996, Section 4.7]),

$$\mathbb{E}_{a,n}(M_0) = a = \mathbb{E}_{a,n}(X_{\mathcal{T}}) - (p-q) \cdot \mathbb{E}_{a,n}(\mathcal{T}) - 2\varepsilon \cdot \mathbb{E}_{a,n}(\eta_{n,m}).$$

Therefore

$$\mathbb{E}_{a,n}(\mathcal{T}) - E_a(\mathcal{T}) = \frac{1}{p-q} \cdot \left[\mathbb{E}_{a,n}(X_{\mathcal{T}}) - E_a(X_{\mathcal{T}}) - 2\varepsilon \cdot \mathbb{E}_{a,n}(\eta_{n,m}) \right].$$

The equilibrium price is defined from

$$\frac{r}{b} \cdot \mathbb{E}_{a,n}(X_{\mathcal{T}}) - c^*_{\text{time}}(a,n) \cdot \mathbb{E}_{a,n}(\eta_{n,m}) - \Lambda \cdot \mathbb{E}_{a,n}(\mathcal{T}) = \frac{r}{b} \cdot E_a(X_{\mathcal{T}}) - \Lambda \cdot E_a(\mathcal{T}).$$

That is,

$$c^*_{\text{time}}(a,n) = \frac{1}{\mathbb{E}_{a,n}(\eta_{n,m})} \left[\frac{r}{b} \cdot \left(\mathbb{E}_{a,n}(X_{\mathcal{T}}) - E_a(X_{\mathcal{T}}) \right) - \Lambda \cdot \left(\mathbb{E}_{a,n}(\mathcal{T}) - E_a(\mathcal{T}) \right) \right],$$

and hence

$$
\begin{aligned}
c^*_{\text{time}}(a,n) &= \frac{1}{\mathbb{E}_{a,n}(\eta_{n,m})} \left[\left(\frac{r}{b} - \frac{\Lambda}{p-q} \right) \cdot \left(\mathbb{E}_{a,n}(X_{\mathcal{T}}) - E_a(X_{\mathcal{T}}) \right) + \frac{2\varepsilon \Lambda}{p-q} \cdot \mathbb{E}_{a,n}(\eta_{n,m}) \right] \\
&= \left(\frac{r}{b} - \frac{\Lambda}{p-q} \right) \cdot \frac{\mathbb{E}_{a,n}(X_{\mathcal{T}}) - E_a(X_{\mathcal{T}})}{\mathbb{E}_{a,n}(\eta_{n,m})} + \frac{2\varepsilon \Lambda}{p-q} \\
&= \left(1 - \frac{\Lambda b}{r(p-q)} \right) \cdot c^*(n) + \frac{2\varepsilon \Lambda}{p-q}.
\end{aligned}
$$

(b) If $p = q$, let

$$M_k = X_k^2 - k - 4\varepsilon \cdot n \cdot \eta_{n,m}(k-1), \quad k \geq 0.$$

As before we convene that $\eta_{n,m}(-1) = 0$. Then $(M_k)_{k \geq 0}$ is martingale with respect to the natural filtration of the Markov chain $(X_k, m_k)_{k \geq 0}$. Hence

$$a^2 = \mathbb{E}_{a,n}(X_{\mathcal{T}}^2) - \mathbb{E}_{a,n}(\mathcal{T}) - 4\varepsilon \cdot n \cdot \mathbb{E}_{a,n}(\eta_{n,m}),$$

and thus

$$\mathbb{E}_{a,n}(\mathcal{T}) - E_a(\mathcal{T}) = [\mathbb{E}_{a,n}(X_{\mathcal{T}}^2) - E_a(X_{\mathcal{T}}^2) - 4\varepsilon \cdot n \cdot \mathbb{E}_{a,n}(\eta_{n,m})].$$

The equilibrium price is defined from the identity

$$\frac{r}{b} \cdot \mathbb{E}_{a,n}(X_{\mathcal{T}}) - c_{\text{time}}^*(n) \cdot \mathbb{E}_{a,n}(\eta_{n,m}) - \Lambda \mathbb{E}_{a,n}(\mathcal{T}) = \frac{r}{b} \cdot E_a(X_{\mathcal{T}}) - \Lambda E_a(\mathcal{T}).$$

That is,

$$c_{\text{time}}^*(n) = \frac{1}{\mathbb{E}_{a,n}(\eta_{n,m})} \left[\frac{r}{b} \cdot (\mathbb{E}_{a,n}(X_{\mathcal{T}}) - E_a(X_{\mathcal{T}})) - \Lambda \cdot (\mathbb{E}_{a,n}(\mathcal{T}) - E_a(\mathcal{T})) \right],$$

and hence

$$c_{\text{time}}^*(n) = \frac{1}{\mathbb{E}_{a,n}(\eta_{n,m})} \left[(r - \Lambda b^2) \cdot (\mathbb{P}_{a,n}(X_{\mathcal{T}}) - P_a(X_{\mathcal{T}})) + 4\varepsilon n \Lambda \cdot \mathbb{E}_{a,n}(\eta_{n,m}) \right]$$

$$= \left(1 - \frac{\Lambda b^2}{r} \right) \cdot c^*(n) + 4\varepsilon n \Lambda,$$

as required. $\square$

7. Chain of stores associated with the 1-excited random walk

One is prompted to study the buyer-seller game described in Section 2 for more complex initial configurations of cookies (store placements). In particular, it is interesting to compare the effect of the *cookie store perturbation* on the underlying random walk in different models. In what follows we focus on finding the equilibrium price for a cookie when $X_0 = 1$ and exactly one cookie is placed at each integer site within the interval $(0, b)$. The corresponding random walk $(X_k)_{k \geq 0}$ is usually referred to as the *1-excited* random walk on $\mathbb{Z}$ (see, for instance, [Antal and Redner 2005; Benjamini and Wilson 2003]). Our main results in this section are stated in Theorems 7.1 and 7.2; see also two remarks concluding the section.

Let $\mathcal{P}_k$ be the probability that the 1-excited random walk starting at $X_0 = 1$ will reach site $k > 0$ before hitting 0. Our results in this section rely on an explicit formula for $\mathcal{P}_k$ and its asymptotic analysis. These quantities are fundamental for the random walk theory. They have been discussed in [Antal and Redner 2005], based on arguments of a different type from ours.

Let $U_c^{\text{we}}(b)$ (here *we* abbreviates "weakly excited") denote the expected earnings of the buyer when the price for a cookie is $c > 0$ and she is using the cookies. We will denote by $c_{\text{we}}^*(b)$ the subgame perfect equilibrium price for a cookie for a buyer performing the 1-excited random walk on $[0, b]$ with absorbing boundaries, starting at $X_0 = 1$. Since $\mathcal{P}_k - \mathcal{P}_{k+1}$ is the probability that the random walk started at $X_0 = 1$ will reach k but never $k + 1$ before the ruin at 0, we have

$$U_c^{\text{we}}(b) = \mathcal{P}_b \cdot [r - c(b - 1)] - \sum_{k=1}^{b-1} (\mathcal{P}_k - \mathcal{P}_{k+1}) \cdot ck.$$

Similarly to (8), we have

$$c^*_{\mathrm{we}}(b) = \frac{r[\mathscr{P}_b - P_1(T_b < T_0)]}{\mathscr{P}_b \cdot (b-1) + \sum_{k=1}^{b-1}(\mathscr{P}_k - \mathscr{P}_{k+1}) \cdot k}. \tag{29}$$

Theorem 7.1. *If $p = q$, we have $\lim_{b \to \infty} bc^*_{\mathrm{we}}(b) = 2r\varepsilon$.*

Proof. We have

$$\mathscr{P}_{k+1} = \mathscr{P}_k \cdot [p + \varepsilon + (q - \varepsilon)P_{k-1}(T_{k+1} < T_0)], \tag{30}$$

which implies, for $p = q$,

$$\mathscr{P}_{k+1} = \mathscr{P}_k \cdot \left[p + \varepsilon + (q - \varepsilon)\frac{k-1}{k+1}\right] = \frac{\mathscr{P}_k \cdot (k + 2\varepsilon)}{k+1}.$$

Thus

$$\mathscr{P}_k = \frac{1}{k!}\prod_{j=1}^{k-1}(j + 2\varepsilon) = \frac{1}{k}\prod_{j=1}^{k-1}\left(1 + \frac{2\varepsilon}{j}\right), \quad k = 1, \ldots, b,$$

with the usual convention that $\prod_{k=1}^{0} a_k = 1$ for any reals a_k. It follows from (29) that

$$c^*_{\mathrm{we}}(b) = \frac{r[\mathscr{P}_b - b^{-1}]}{\mathscr{P}_b \cdot (b-1) + \sum_{k=1}^{b-1}(\mathscr{P}_k - \mathscr{P}_{k+1}) \cdot k}.$$

Observe that

$$(\mathscr{P}_k - \mathscr{P}_{k+1}) \cdot k = \mathscr{P}_k \frac{k(1 - 2\varepsilon)}{k+1}. \tag{31}$$

We will next show that

$$\lim_{n \to \infty} n^{1-2\varepsilon}\mathscr{P}_n = c_\varepsilon > 0 \quad \text{for some constant } c_\varepsilon > 0. \tag{32}$$

Let $f_n = n^{1-2\varepsilon}\mathscr{P}_n$. Then

$$\frac{f_{n+1}}{f_n} = \frac{(n+1)^{1-2\varepsilon}(n+2\varepsilon)}{(n+1)n^{1-2\varepsilon}} = \frac{(n+1)^{-2\varepsilon}}{n^{-2\varepsilon}}\frac{n+2\varepsilon}{n} = \left(1 + \frac{1}{n}\right)^{-2\varepsilon} \cdot \left(1 + \frac{2\varepsilon}{n}\right)$$

$$< \left(1 + \frac{2\varepsilon}{n}\right)^{-1} \cdot \left(1 + \frac{2\varepsilon}{n}\right) = 1.$$

Therefore, f_n is an increasing sequence. On the other hand, using convexity of the function $g(x) = 1/x$ and the inequality $1 + x \le e^x$, $x \in \mathbb{R}$, we obtain

$$f_n = n^{1-2\varepsilon}\mathscr{P}_n = n^{-2\varepsilon}\prod_{j=1}^{n-1}\left(1 + \frac{2\varepsilon}{j}\right) \le n^{-2\varepsilon}\exp\left(\sum_{j=1}^{n-1} 2\varepsilon j^{-1}\right)$$

$$< n^{-2\varepsilon}\exp\left(2\varepsilon + 2\varepsilon\int_1^{n-1} x^{-1}\,dx\right) < e^{2\varepsilon} < \infty.$$

Therefore, f_n converges to a finite nonzero limit when n approaches infinity. Furthermore, according to (32), f_n is a regularly varying at infinity sequence of index $-(1-2\varepsilon)$ [Bojanic and Seneta 1973]. This implies [ibid., Theorem 6]

$$\lim_{b\to\infty} (b^2 f_b)^{-1} \sum_{k=1}^{b} k^2 f_k (k+1)^{-1} = (2\varepsilon)^{-1}.$$

This observation along with (31) imply

$$\begin{aligned}
\lim_{b\to\infty} b\cdot c^*_{\text{we}}(b) &= \lim_{b\to\infty} \frac{br(\mathcal{P}_b - b^{-1})}{\mathcal{P}_b\cdot(b-1)+\sum_{k=1}^{b-1}(\mathcal{P}_k-\mathcal{P}_{k+1})\cdot k} \\
&= \lim_{b\to\infty} \frac{br\mathcal{P}_b}{\mathcal{P}_b\cdot(b-1)+(2\varepsilon)^{-1}(\mathcal{P}_{b-1}-\mathcal{P}_b)\cdot(b-1)b} \\
&= \lim_{b\to\infty} \frac{br\mathcal{P}_b}{\mathcal{P}_b\cdot(b-1)+(2\varepsilon)^{-1}\mathcal{P}_{b-1}(b-1)(1-2\varepsilon)} = \frac{r}{1+(2\varepsilon)^{-1}(1-2\varepsilon)} \\
&= 2\varepsilon r.
\end{aligned}$$

The proof of the theorem is completed. $\qquad\square$

For $p \neq q$, recurrence relation (30) implies

$$\begin{aligned}
\mathcal{P}_{k+1} &= \mathcal{P}_k \cdot \left[p + \varepsilon + (q-\varepsilon)\frac{\rho^{k-1}-1}{\rho^{k+1}-1} \right] = \mathcal{P}_k \frac{\rho^k - 1 + \varepsilon(\rho^{k+1}-\rho^{k-1})}{\rho^{k+1}-1} \\
&= \mathcal{P}_k \left(1 + \varepsilon \frac{\rho^{k+1}-\rho^{k-1}}{\rho^k-1} \right) \frac{\rho^k-1}{\rho^{k+1}-1}.
\end{aligned}$$

Thus $\mathcal{P}_1 = 1$ and

$$\mathcal{P}_k = \frac{\rho-1}{\rho^k-1} \prod_{j=1}^{k-1} \left(1 + \varepsilon \frac{\rho^{j+1}-\rho^{j-1}}{\rho^j-1} \right), \quad k = 2,\ldots,b.$$

In this case

$$c^*_{\text{we}}(b) = \frac{r\left(\mathcal{P}_b - \frac{\rho-1}{\rho^b-1}\right)}{\mathcal{P}_b\cdot(b-1)+\sum_{k=1}^{b-1}(\mathcal{P}_k-\mathcal{P}_{k+1})\cdot k}. \tag{33}$$

Observe that

$$\begin{aligned}
(\mathcal{P}_k - \mathcal{P}_{k+1}) &= \mathcal{P}_k \left(1 - \frac{\rho^k-1+\varepsilon(\rho^{k+1}-\rho^{k-1})}{\rho^{k+1}-1} \right) \\
&= \mathcal{P}_k \left(\frac{\rho^{k+1}-\rho^k-\varepsilon(\rho^{k+1}-\rho^{k-1})}{\rho^{k+1}-1} \right) \\
&= \mathcal{P}_k \cdot \rho^{k-1} \left(\frac{\rho(\rho-1)-\varepsilon(\rho^2-1)}{\rho^{k+1}-1} \right). \tag{34}
\end{aligned}$$

It follows from (33) that

$$c_{\text{we}}^*(b) \leq \frac{r\mathscr{P}_b}{\mathscr{P}_b \cdot (b-1)} = \frac{r}{b-1}.$$

The following theorem shows that this bound is asymptotically tight for $\rho < 1$, regardless the value of ε.

Theorem 7.2.   (a) *If $\rho > 1$, we have*

$$\lim_{b \to \infty} \left(\frac{\rho}{1+\varepsilon(\rho-\rho^{-1})} \right)^b c_{\text{we}}^*(b) = c_\varepsilon \quad \text{for some constant } c_\varepsilon \in (0, \infty).$$

(b) *If $\rho < 1$, we have* $\lim_{b \to \infty} b c_{\text{we}}^*(b) = r$.

Proof. (a) Assume that $\rho > 1$. We will first show that

$$\lim_{n \to \infty} \left(\frac{\rho}{1+\varepsilon(\rho-\rho^{-1})} \right)^n \mathscr{P}_n = \tilde{c}_\varepsilon \quad \text{for some constant } \tilde{c}_\varepsilon \in (0, \infty). \qquad (35)$$

Notice that $\frac{\rho}{1+\varepsilon(\rho-\rho^{-1})} > 1$ because $\varepsilon < q$. Let

$$f_n = \left(\frac{\rho}{1+\varepsilon(\rho-\rho^{-1})} \right)^n \mathscr{P}_n.$$

Then

$$\frac{f_{n+1}}{f_n} = \frac{\rho}{1+\varepsilon(\rho-\rho^{-1})} \cdot \frac{\rho^n - 1}{\rho^{n+1} - 1} \cdot \left(1 + \varepsilon \frac{\rho^{n+1} - \rho^{n-1}}{\rho^n - 1} \right)$$

$$= \frac{\rho}{1+\varepsilon(\rho-\rho^{-1})} \cdot \frac{\rho^n - 1 + \varepsilon(\rho^{n+1} - \rho^{n-1})}{\rho^{n+1} - 1}$$

$$= \frac{\rho^{n+1} - \rho + \varepsilon(\rho^{n+2} - \rho^n)}{\rho^{n+1} - 1 + \varepsilon(\rho^{n+2} - \rho^n - \rho + \rho^{-1})} < 1.$$

To verify the inequality in the last line above write

$$\rho^{n+1} - \rho + \varepsilon(\rho^{n+2} - \rho^n) < \rho^{n+1} - 1 + \varepsilon(\rho^{n+2} - \rho^n - \rho + \rho^{-1})$$

$$\iff \varepsilon(\rho - \rho^{-1}) < \rho - 1 \iff \varepsilon(\rho + 1) < \rho.$$

The last inequality is true because $\varepsilon < q$. Thus, we have proved that f_n is a decreasing sequence. On the other hand, since

$$\rho^{n-1} \frac{\rho - 1}{\rho^n - 1} > \frac{\rho - 1}{\rho},$$

we obtain

$$f_n = \left(\frac{\rho}{1 + \varepsilon(\rho - \rho^{-1})}\right)^{n-1} \mathcal{P}_n$$

$$\geq \frac{\rho - 1}{\rho} \cdot \left(\frac{1}{1 + \varepsilon(\rho - \rho^{-1})}\right)^{n-1} \prod_{j=1}^{n-1}\left(1 + \varepsilon\frac{\rho^{j+1} - \rho^{j-1}}{\rho^j - 1}\right)$$

$$\geq \frac{\rho - 1}{\rho} \cdot \left(\frac{1}{1 + \varepsilon(\rho - \rho^{-1})}\right)^{n-1} \prod_{j=1}^{n-1}(1 + \varepsilon(\rho - \rho^{-1})) > \frac{\rho - 1}{\rho} > 0.$$

Therefore, f_n is a bounded away from zero decreasing sequence, and hence $\lim_{n \to \infty} f_n$ exists and is strictly positive and finite. Notice that

$$\frac{\rho}{1 + \varepsilon(\rho - \rho^{-1})} < \rho,$$

and hence

$$\lim_{n \to \infty} \mathcal{P}_b\left(\frac{\rho - 1}{\rho^b - 1}\right)^{-1} = \infty.$$

Therefore, due to (34) and (35), the following limit exists and is strictly positive and finite:

$$\lim_{b \to \infty} \left(\frac{\rho}{1 + \varepsilon(\rho - \rho^{-1})}\right)^b c_{\text{we}}^*(b) = \lim_{b \to \infty} \frac{\left(\frac{\rho}{1 + \varepsilon(\rho - \rho^{-1})}\right)^b r \mathcal{P}_b}{\mathcal{P}_b \cdot (b - 1) + \sum_{k=1}^{b-1}(\mathcal{P}_k - \mathcal{P}_{k+1}) \cdot k}$$

$$= \frac{r\tilde{c}_\varepsilon}{\sum_{k=1}^{\infty}(\mathcal{P}_k - \mathcal{P}_{k+1}) \cdot k} := c_\varepsilon \in (0, \infty).$$

(b) We now turn to the case $\rho < 1$. In virtue of (33) and (34) it suffices to show that

$$\lim_{n \to \infty} \mathcal{P}_n = \lim_{n \to \infty} (1 - \rho) \prod_{j=1}^{n-1}\left(1 + \varepsilon\frac{\rho^{j+1} - \rho^{j-1}}{\rho^j - 1}\right) = \hat{c}_\varepsilon$$

for some constant $\hat{c}_\varepsilon \in (0, \infty)$. Let

$$f_n = \prod_{j=1}^{n-1}\left(1 + \varepsilon\frac{\rho^{j+1} - \rho^{j-1}}{\rho^j - 1}\right).$$

Then f_n is an increasing sequence. On the other hand,

$$f_n = \prod_{j=1}^{n-1}\left(1 + \varepsilon\frac{\rho^{j+1} - \rho^{j-1}}{\rho^j - 1}\right) \leq \prod_{j=1}^{n-1}\left(1 + \varepsilon\frac{\rho^{j+1} - \rho^{j-1}}{\rho - 1}\right)$$

$$\leq \exp\left(\sum_{j=1}^{\infty} \varepsilon\frac{\rho^{j+1} - \rho^{j-1}}{\rho - 1}\right) \exp\left(\varepsilon\frac{1 + \rho}{1 - \rho}\right) < \infty.$$

Therefore, f_n is a bounded and increasing sequence, and hence $\lim_{n\to\infty}\mathcal{P}_n = (1-\rho)\lim_{n\to\infty}f_n$ exists and is strictly positive and finite. This completes the proof of the theorem. $\qquad\square$

Remark 7.3. We notice that Theorem 7.1 and Theorem 7.2 can be alternatively stated as follows. We will write $a_n \sim b_n$ when $\lim_{n\to\infty}a_n/b_n = 1$ for two sequences of real numbers $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$.

(a) If $p = q$ and r depends on b in such a way that $r(b) \sim cb$ for some constant $c \in (0,\infty)$, then $\lim_{b\to\infty}c^*_{\text{we}}(b) = 2c\varepsilon$.

(b) If $\rho > 1$ and r depends on b in such a way that

$$r(b) \sim c\left(\frac{\rho}{1+\varepsilon(\rho-\rho^{-1})}\right)^b$$

for some constant $c \in (0,\infty)$, then $\lim_{b\to\infty}c^*_{\text{we}}(b) = c_\varepsilon$ for some constant $c_\varepsilon \in (0,\infty)$.

(c) If $\rho < 1$ and r depends on b in such a way that $r(b) \sim cb$ for some constant $c \in (0,\infty)$, then $\lim_{b\to\infty}c^*_{\text{we}}(b) = c$.

Remark 7.4. Let $V^*_{\text{we}}(b)$ denote the expected revenue of the seller at the equilibrium. Then $V^*_{\text{we}}(b) = r \cdot [\mathcal{P}_b - b^{-1}] \sim r\mathcal{P}_b$ as b goes to infinity. Thus the asymptotic for $\mathcal{P}_b$ found in the course of the proof of Theorems 7.1 and 7.2 (see also [Antal and Redner 2005] for a heuristic derivation) yields the asymptotic for $V^*_{\text{we}}(b)$. More precisely, for some strictly positive constants c_ε, $\tilde{c}_\varepsilon$, and $\hat{c}_\varepsilon$ we have, as b goes to infinity:

(a) If $p = q$, then $V^*_{\text{we}}(b) \sim c_\varepsilon b^{-(1-2\varepsilon)}$.

(b) If $\rho > 1$, then $V^*_{\text{we}}(b) \sim \tilde{c}_\varepsilon\left(\frac{\rho}{1+\varepsilon(\rho-\rho^{-1})}\right)^{-b}$.

(c) If $\rho < 1$, then $V^*_{\text{we}}(b) \sim \hat{c}_\varepsilon$.

8. Conclusion

We explored a simple game-theoretic modification of the gambler's ruin problem. The underlying random walk is defined through a single-point perturbation of the transition probabilities of the regular nearest-neighbor random walk on $\mathbb{Z}$, either recurrent or transient. The perturbation is the same as the one in the excited (cookie) random walks model, except being localized to a single point. Informally, the deformation of the transition kernel can be described as a store that provides an instant increase in probability in the positive direction when the buyer visits the store. The price of a cookie is determined in the game (negotiation) between the buyer (walker) and the seller (store's owner). The equilibrium price can vary,

depending on the store's location. The seller chooses the location to maximize her expected revenue. The goal of the buyer in the game is to maximize her expected earning which is expressed in terms of a utility function. An analytical equation for the equilibrium price, given the starting point of the walker and the store's location, is derived for several interesting choices of the utility function, including risk-neutral behavior model, risk-averse behavior model, and a model including an opportunity cost represented by time spent in the game. The difference between the equilibrium price policies associated with different utility functions is quite intuitive. The equilibrium price of the cookie has a nice scaling property when the range of the interval approaches infinity. Thus the price is a natural characteristic capturing the global effect of the "cookie store perturbation" on the regular random walk. In fact, the structure of the equilibrium price is closely related to the structure of exit probabilities (and local times) of the underlying (both perturbed and not perturbed) random walks. For comparison, we include similar asymptotic results for 1-excited random walk in our analysis. In principle, the spatial distribution of the equilibrium price allows us to recover the optimal store location. The optimal store placement coincides with the buyer's starting point for the basic model of a risk-neutral buyer, whereas in other cases it can be determined with the help of numerical analysis. In a future work we consider continuous-time versions of the problems studied in this paper by replacing the nearest-neighbor random walk with a drifted Brownian motion. In a paper in preparation we enrich the game-theoretic component of the basic game by including a third player, modeling both duopoly competition and a state regulation of the market.

Acknowledgment

References

[Anderson et al. 1992] S. P. Anderson, A. de Palma, and J.-F. Thisse, *Discrete choice theory of product differentiation*, MIT Press, Cambridge, MA, 1992. MR 94b:90012 Zbl 0857.90018

[Antal and Redner 2005] T. Antal and S. Redner, "The excited random walk in one dimension", *J. Phys. A* **38**:12 (2005), 2555–2577. MR 2005k:82026 Zbl 1113.82024

[Bell 1988] D. E. Bell, "One-switch utility functions and a measure of risk", *Management Sci.* **34**:12 (1988), 1416–1424. MR 89j:90031 Zbl 0665.90006

[Bell and Fishburn 2001] D. E. Bell and P. C. Fishburn, "Strong one-switch utility", *Management Sci.* **47** (2001), 601–604.

[Benjamini and Wilson 2003] I. Benjamini and D. B. Wilson, "Excited random walk", *Electron. Comm. Probab.* **8** (2003), 86–92. MR 2004b:60120 Zbl 1060.60043

[Bojanic and Seneta 1973] R. Bojanic and E. Seneta, "A unified theory of regularly varying sequences", *Math. Z.* **134** (1973), 91–106. MR 48 #11407 Zbl 0256.40002

[Davis 1999] B. Davis, "Brownian motion and random walk perturbed at extrema", *Probab. Theory Related Fields* **113**:4 (1999), 501–518. MR 2001k:60030 Zbl 0930.60041

[Durrett 1996] R. Durrett, *Probability: Theory and examples*, 2nd ed., Duxbury Press, Belmont, CA, 1996. MR 98m:60001 Zbl 0709.60002

[El-Shehawey 2009] M. A. El-Shehawey, "On the gambler's ruin problem for a finite Markov chain", *Statist. Probab. Lett.* **79**:14 (2009), 1590–1595. MR 2010i:60133 Zbl 1172.60311

[Gibbons 1992] R. Gibbons, *Game theory for applied economists*, Princeton University Press, 1992.

[Hotelling 1929] H. Hotelling, "Stability in competition", *Econ. J.* **39** (1929), 41–57.

[Menshikov et al. 2012] M. Menshikov, S. Popov, A. Ramírez, and M. Vachkovskaia, "On a general many-dimensional excited random walk", *Annals of Probability* **40**:5 (2012), 2106–2130. MR 3025712 Zbl 06111052

[O'Donoghue and Rabin 1999] E. O'Donoghue and M. Rabin, "Doing it now or later", *Amer. Econ. Rev.* **89** (1999), 103–124.

[Raimond and Schapira 2010] O. Raimond and B. Schapira, "Random walks with occasionally modified transition probabilities", *Illinois J. Math.* **54**:4 (2010), 1213–1238. MR 2981846 Zbl 06122093

[Zerner 2005] M. P. W. Zerner, "Multi-excited random walks on integers", *Probab. Theory Related Fields* **133**:1 (2005), 98–122. MR 2006k:60178 Zbl 1076.60088

kuejai@gmail.com   *Department of Mathematics, Iowa State University, Ames, IA 50011, United States*

tpluta@ncsu.edu   *Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, United States*

rastegar@iastate.edu   *Department of Mathematics, Iowa State University, Ames, IA 50011, United States*

roiterst@iastate.edu   *Department of Mathematics, Iowa State University, Ames, IA 50011, United States*

mtemba@umd.edu   *Department of Mathematics, University of Maryland, College Park, MD 20742-4015, United States*

cvidden@iastate.edu   *Department of Mathematics, Iowa State University, Ames, IA 50011, United States*

brian.george.wu@gmail.com *Department of Mathematics, Bowdoin College, Brunswick, ME 04011, United States*

msp

Decomposing induced characters
of the centralizer of an n-cycle
in the symmetric group on $2n$ elements

Joseph Ricci

(Communicated by Nigel Boston)

We give explicit multiplicities and formulas for multiplicities of the characters appearing in the decomposition of the induced character $\mathrm{Ind}_{C_{S_{2n}}(\sigma)}^{S_{2n}} 1_C$, where σ is an n-cycle, $C_{S_{2n}}(\sigma)$ is the centralizer of σ in S_{2n}, and 1_C is the trivial character on $C_{S_{2n}}(\sigma)$.

1. Introduction

Throughout this paper we work only over the complex numbers, dealing with $\mathbb{C}S_n$ characters, where S_n is the symmetric group on n elements. Let $\sigma \in S_n$. In a natural way, by fixing $n+1, \ldots, 2n$, we can regard σ as an element of S_{2n} as well. Let $C := C_{S_{2n}}(\sigma)$ be the centralizer of σ in S_{2n}. Let ψ be any linear character of C. Hemmer [2011] showed that for $m \geq n$ the induced character $\mathrm{Ind}_C^{S_m} \psi$ becomes representation stable for $m = 2n$. Therefore, these induced characters arise naturally when studying braid group cohomology. (For more on representation stability and braid group cohomology, see [Church and Farb 2010].) It was proposed that in general the decomposition of the induced character $\mathrm{Ind}_C^{S_{2n}} \psi$ into irreducible characters of S_{2n} was an open problem.

However, the case when $\sigma = (1\ 2\ \cdots\ n)$ was studied in [Jöllenbeck and Schocker 2000; Kraśkiewicz and Weyman 2001]. In this case, $C_{S_n}(\sigma) = \langle \sigma \rangle$. Then the linear characters of C are precisely the irreducible characters, which are indexed by the numbers $k = 0, 1, \ldots, n-1$ and take σ to $e^{\frac{2\pi i k}{n}}$. It was shown that, for an irreducible character χ^λ of S_n, the multiplicity of χ^λ in the decomposition of $\mathrm{Ind}_{\langle \sigma \rangle}^{S_n} \psi_k$ is equal to the number of standard Young tableaux of shape λ with major index congruent to $k \mod n$. Once this is computed, one can use the Littlewood–Richardson rule or the branching rule to induce the resulting characters up to S_{2n}. So, in theory, the decomposition of $\mathrm{Ind}_C^{S_{2n}} \psi_k$ is known; however, no explicit formula is available in general.

In this paper we will deal with the case when σ is an n-cycle of S_n and $\psi_k = 1_C$ (i.e., $k = 0$), the trivial character. We present a partial result toward an explicit formula as well as a formula for the multiplicities of certain irreducible $\mathbb{C}S_{2n}$ characters appearing in the decomposition.

2. Preliminaries

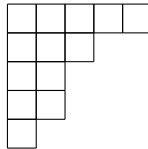### *Partitions and Young diagrams.*

Definition 2.1. We say that $\lambda = (\lambda_1, \ldots, \lambda_r)$ is a *partition* of n, written $\lambda \vdash n$, if $\lambda_i \geq \lambda_{i+1} \geq 0$ for each $\lambda_i \in \mathbb{Z}$ and $\lambda_1 + \cdots + \lambda_r = n$. We say each λ_i is a *part* of λ.

Definition 2.2. Let $\lambda = (\lambda_1, \ldots, \lambda_r) \vdash n$. The *Young diagram*, $[\lambda]$, of λ is the set

$$[\lambda] = \{(i, j) \in \mathbb{N} \times \mathbb{N} \mid j \leq \lambda_i\}.$$

We say each $(i, j) \in [\lambda]$ is a *node* of $[\lambda]$.

If $\lambda \vdash n$, we represent $[\lambda]$ by an array of boxes. As an example, consider the partition $\lambda = (5, 3, 2, 2, 1) \vdash 13$. Then we visualize $[\lambda]$ as

where the upper left box is defined to be the ordered pair $(1, 1)$, the upper right is $(1, 5)$, the lower left is $(5, 1)$, just like the entries of a matrix.

We will often drop the bracket notation and use λ and $[\lambda]$ interchangeably, though it will be clear by context to which we are referring. If λ_i is a part of $\lambda \vdash n$, then λ/λ_i is the partition of $n - \lambda_i$ formed by deleting λ_i from λ. So $(5, 3, 2, 2, 1)/\lambda_2 = (5, 2, 2, 1)$. If $b = (i, j)$ is a node in the Young diagram of λ, we will write $b \in \lambda$. Suppose $\mu = (3, 2, 1)$. Returning to our previous example, it is easy to see that each node $b \in \mu$ is also a node of λ. We will denote this in the obvious way, $\mu \subseteq \lambda$. With this idea in mind, we make a definition.

Definition 2.3. Let λ and μ be partitions such that $\mu \subseteq \lambda$. Then the *skew diagram* λ/μ is the set of nodes

$$\xi = \lambda/\mu = \{b \in \lambda \mid b \notin \mu\}.$$

In the case of our example, the skew diagram λ/μ would be this:
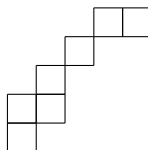
One important aspect of Young diagrams that will be of great important in this paper are rim hooks.

Definition 2.4. For a skew diagram ξ, we say the unique node (i_0, j_0) such that $i_0 \leq i$ and $j_0 \geq j$ for all $(i, j) \in \xi$ is the *top node* of ξ.

Definition 2.5. A *rim hook* is a skew diagram ξ such that if (i, j) is not the top node of ξ then either $(i - 1, j) \in \xi$ or $(i, j + 1) \in \xi$, but not both.

We will say a rim k-hook or simply a k-hook is a rim hook consisting of k nodes. We will say that a partition λ *has* a k-hook if it is possible to remove a k-hook from λ and have the resulting diagram be the Young diagram of some partition λ'. To each rim hook ξ is assigned the leg length of ξ.

Definition 2.6. Let ξ be a rim hook. The *leg length of* ξ, denoted by $ll(\xi)$, is

$$ll(\xi) = (\text{the number of rows in } \xi) - 1.$$

Once again returning to our example where $\lambda = (5, 3, 2, 2, 1)$, we see that λ has three rim 4-hooks:
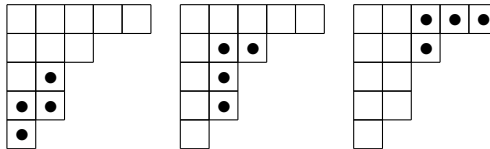
In the first and third cases, the 4-hooks have leg length 2, while in the second case the 4-hook has leg length 1. One can also see that λ does not have any rim 5-hooks, since it is not possible to remove a 5-hook from λ and have the resulting diagram be the Young diagram of a partition.

Character theory of the symmetric group. The basics of representation and character theory will be assumed, and can be found in [James and Liebeck 2001]. It is well known [Sagan 2001, 2.3.4, 2.4.4] that there is a one-to-one correspondence between the set of partitions of n and the set of irreducible characters of S_n. For example, $\chi^{(n)}$ corresponds to the trivial character, $\chi^{(n-1,1)}$ corresponds to the number of fixed points minus one, and $\chi^{(1^n)}$ corresponds to the sign character. Also, the conjugacy classes of S_n have a natural correspondence to the partitions of n. If $\tau \in S_n$ is of cycle type λ, $\lambda \vdash n$, then we will denote the conjugacy class of τ by K_λ. Let $\lambda, \mu \vdash n$. Suppose one wants to evaluate the character χ^λ on the conjugacy class K_μ, which we will denote by χ_μ^λ. The following theorem, known as the *Murnaghan–Nakayama rule*, allows one to recursively compute χ_μ^λ:

Theorem 2.7 [Sagan 2001, 4.10.2]. *Let* $\mu = (\mu_1, \ldots, \mu_s)$ *and assume* $\mu, \lambda \vdash n$. *Then*

$$\chi_\mu^\lambda = \sum_\xi (-1)^{ll(\xi)} \chi_{\mu/\mu_1}^{\lambda/\xi},$$

where the sum is taken over all rim hooks ξ *of* λ *containing* μ_1 *nodes.*

Now, in a natural way, one can think of S_{n-1} as a subgroup of S_n. Suppose χ^λ is the character of S_n corresponding to λ and χ^μ is the character of S_{n-1} corresponding to μ. Then one can easily compute the restricted character $\chi^\lambda \downarrow_{S_{n-1}}$ and the induced character $\mathrm{Ind}_{S_{n-1}}^{S_n} \chi^\mu$ using the branching rule.

Definition 2.8. Let $\lambda \vdash n$. We say an *inner corner* of $[\lambda]$ is a node $(i, j) \in [\lambda]$ such that $[\lambda] - \{(i, j)\}$ is the Young diagram of some partition of $n - 1$. We denote any such partition by λ^-. We say an *outer corner* is a node $(i, j) \notin [\lambda]$ such that $[\lambda] \cup \{(i, j)\}$ is the Young diagram of some partition of $n + 1$. We denote any such partition by λ^+.

Theorem 2.9 (branching rule [Sagan 2001, 2.8.3]). *Let* $\mu \vdash n - 1$, $\lambda \vdash n$. *Then*

$$\chi^\lambda \downarrow_{S_{n-1}} = \sum_{\lambda^-} \chi^{\lambda^-} \quad and \quad \mathrm{Ind}_{S_{n-1}}^{S_n} \chi^\mu = \sum_{\mu^+} \chi^{\mu^+}.$$

As an example, suppose $\lambda = (3, 3, 2)$ and $\mu = (5, 2)$. Using Theorem 2.9 we calculate

$$\chi^{(3,3,2)} \downarrow_{S_7} = \chi^{(3,2,2)} + \chi^{(3,3,1)},$$
$$\mathrm{Ind}_{S_7}^{S_8} \chi^{(5,2)} = \chi^{(6,2)} + \chi^{(5,3)} + \chi^{(5,2,1)}.$$

3. The decomposition of ϕ

Some preliminary results. Recall that in the introduction we defined $C := C_{S_{2n}}(\sigma)$, with $\sigma = (1\ 2\ \cdots\ n)$. One can compute that $C \cong \langle \sigma \rangle \times S_n$ [Dummit and Foote 2004, 4.3]. Keeping this in mind we have the following notation:

Notation. For $\tau \in C$, we will write $\tau = (\sigma^k, \pi)$ for $k \in \mathbb{Z}$ and $\pi \in S_n$.

Also, if $\lambda = (\lambda_1, \ldots, \lambda_r) \vdash n$ then

$$(n, \lambda) := (n, \lambda_1, \ldots, \lambda_r) \vdash 2n \quad and \quad (\lambda, 1^n) := (\lambda_1, \ldots, \lambda_r, 1^n) \vdash 2n.$$

Notation. When evaluating any character χ on the conjugacy class of S_{2n} corresponding to (n, λ) or $(\lambda, 1^n)$, we will write $\chi_{(n,\lambda)}$ and $\chi_{(\lambda,1^n)}$, respectively.

For the remainder of this paper, we will write $\phi = \mathrm{Ind}_C^{S_{2n}} 1$.

Proposition 3.1. *Let* $n \geq 1$. *Let* $\chi^{(2n)}$ *be the irreducible character of* S_{2n} *corresponding to the partition* $(2n)$. *Then*

$$\langle \phi, \chi^{(2n)} \rangle_{S_{2n}} = 1.$$

Proof. Using Frobenius reciprocity, we have

$$\langle \phi, \chi^{(2n)} \rangle_{S_{2n}} = \langle 1_C, \chi^{(2n)} \downarrow_C \rangle_C.$$

But since $\chi^{(2n)}$ is the trivial character, $\chi^{(2n)} \downarrow_C = 1_C$, so we have

$$\langle \phi, \chi^{(2n)} \rangle_{S_{2n}} = 1. \qquad \square$$

Proposition 3.2. *Let $n \geq 2$. Let $\chi^{(2n-1,1)}$ be the irreducible character of S_{2n} corresponding to $(2n - 1, 1)$. Then*

$$\langle \phi, \chi^{(2n-1,1)} \rangle_{S_{2n}} = 1.$$

Proof. First note that this character records the number points fixed by a permutation and subtracts 1. Using Frobenius reciprocity, we expand the inner product as follows:

$$\langle \phi, \chi^{(2n-1,1)} \rangle_{S_{2n}} = \langle 1_C, \chi^{(2n-1,1)} \downarrow_C \rangle_C = \frac{1}{nn!} \sum_{\tau \in C} \chi^{(2n-1,1)}(\tau). \qquad (3\text{-}1)$$

By remarks made at the beginning of this section, the last term in (3-1) becomes

$$\frac{1}{nn!} \sum_{k=0}^{n-1} \sum_{\pi \in S_n} \chi^{(2n-1,1)}((\sigma^k, \pi)).$$

When $k = 0$, $(\sigma^k, \pi) = (1, \pi)$ and $(1, \pi)$ fixes $n + \chi^{(n-1,1)}(\pi) + 1$ points. When $k \neq 0$, (σ^k, π) fixes $\chi^{(n-1,1)}(\pi) + 1$ points, giving

$$\frac{1}{nn!} \sum_{k=0}^{n-1} \sum_{\pi \in S_n} \chi^{(2n-1,1)}((\sigma^k, \pi))$$

$$= \frac{1}{nn!} \left(\sum_{\pi \in S_n} \left(n + \chi^{(n-1,1)}(\pi) \right) + (n - 1) \sum_{\pi \in S_n} \chi^{(n-1,1)}(\pi) \right)$$

$$= \frac{1}{nn!} \left(\sum_{\pi \in S_n} n + \sum_{\pi \in S_n} \chi^{(n-1,1)}(\pi) + (n - 1) \sum_{\pi \in S_n} \chi^{(n-1,1)}(\pi) \right)$$

$$= \frac{1}{nn!} \left(nn! + nn! \langle \chi^{(n)}, \chi^{(n-1,1)} \rangle_{S_n} \right). \qquad (3\text{-}2)$$

But since both $\chi^{(n)}$ and $\chi^{(n-1,1)}$ are irreducible, their inner product is 0. So (3-2) becomes

$$\frac{1}{nn!} nn! = 1. \qquad \square$$

Proposition 3.3. *Let $n \geq 2$. Let $\chi^{(n,n)}$ be the irreducible character of S_{2n} corresponding to (n, n). Then*

$$\langle \phi, \chi^{(n,n)} \rangle_{S_{2n}} = 1.$$

Proof. Throughout, let $d_k = \gcd(n, k)$. Using Frobenius reciprocity, we write

$$\langle \phi, \chi^{(n,n)} \rangle_{S_{2n}} = \langle 1_C, \chi^{(n,n)} \downarrow_C \rangle_C = \frac{1}{nn!} \sum_{k=0}^{n-1} \sum_{\pi \in S_n} \chi^{(n,n)}((\sigma^k, \pi)).$$

We break the sum up into three pieces: one for $k = 0$, one for $d_k = 1$ (of which there are $\varphi(n)$ such k, where φ denotes Euler's totient function) and one for $d_k \neq 1$:

$$\langle \phi, \chi^{(n,n)} \rangle_{S_{2n}}$$
$$= \frac{1}{nn!} \left(\sum_{\pi \in S_n} \chi^{(n,n)}((1, \pi)) + \varphi(n) \sum_{\pi \in S_n} \chi^{(n,n)}((\sigma, \pi)) + \sum_{\substack{1 < k < n \\ d_k \neq 1}} \sum_{\pi \in S_n} \chi^{(n,n)}((\sigma^k, \pi)) \right).$$

In order to use Theorem 2.7, we sum over all partitions of n and rewrite the sum as

$$\langle \phi, \chi^{(n,n)} \rangle_{S_{2n}} = \frac{1}{nn!} \left(n! \langle \chi^{(n)}, \chi^{(n,n)} \downarrow_{S_n} \rangle_{S_n} + \varphi(n) \sum_{\lambda \vdash n} \chi^{(n,n)}_{(n,\lambda)} |K_\lambda| \right.$$
$$\left. + \sum_{\substack{1 < k < n \\ d_k \neq 1}} \sum_{\lambda \vdash n} \chi^{(n,n)}_{((\frac{n}{d_k})^{d_k}, \lambda)} |K_\lambda| \right). \quad (3\text{-}3)$$

By Theorem 2.9, we write

$$\chi^{(n,n)} \downarrow_{S_n} = \chi^{(n)} + \sum_{\substack{\lambda \vdash n \\ \lambda \neq (n)}} a_\lambda \chi^\lambda$$

where $a_\lambda \in \{0, 1, 2, \dots\}$. Then, by linearity, we have

$$\langle \chi^{(n)}, \chi^{(n,n)} \downarrow_{S_n} \rangle_{S_n} = \langle \chi^{(n)}, \chi^{(n)} \rangle_{S_n} + \sum_{\substack{\lambda \vdash n \\ \lambda \neq (n)}} a_\lambda \langle \chi^{(n)}, \chi^\lambda \rangle_{S_n}$$
$$= \langle \chi^{(n)}, \chi^{(n)} \rangle_{S_n} = 1, \quad (3\text{-}4)$$

since all the χ^λ are irreducible. Using Theorem 2.7,

$$\chi^{(n,n)}_{(n,\lambda)} = \chi^{(n)}_\lambda - \chi^{(n-1,1)}_\lambda$$

so that

$$\sum_{\lambda \vdash n} \chi^{(n,n)}_{(n,\lambda)} |K_\lambda| = \sum_{\lambda \vdash n} (\chi^{(n)}_\lambda - \chi^{(n-1,1)}_\lambda) |K_\lambda|$$
$$= \sum_{\lambda \vdash n} \chi^{(n)}_\lambda |K_\lambda| - \sum_{\lambda \vdash n} \chi^{(n-1,1)}_\lambda |K_\lambda|$$
$$= n! \langle \chi^{(n)}, \chi^{(n)} \rangle_{S_n} - n! \langle \chi^{(n)}, \chi^{(n-1,1)} \rangle_{S_n} = n!. \quad (3\text{-}5)$$

Now let $d_k \neq 1$, for some k. Again with Theorem 2.7, we write

$$\chi^{(n,n)}_{((\frac{n}{d_k})^{d_k},\lambda)} = \chi^{(n)}_\lambda + \sum_{\substack{\mu \vdash n \\ \mu \neq (n)}} c_\mu \chi^\mu_\lambda$$

where $c_\lambda \in \mathbb{Z}$. Then

$$\sum_{\lambda \vdash n} \chi^{(n,n)}_{((\frac{n}{d_k})^{d_k},\lambda)} |K_\lambda| = \sum_{\lambda \vdash n} \chi^{(n)}_\lambda |K_\lambda| + \sum_{\lambda \vdash n} \sum_{\substack{\mu \vdash n \\ \mu \neq (n)}} c_\mu \chi^\mu_\lambda |K_\lambda|$$

$$= n! \langle \chi^{(n)}, \chi^{(n)} \rangle_{S_n} + \sum_{\substack{\mu \vdash n \\ \mu \neq (n)}} n! \, c_\mu \langle \chi^{(n)}, \chi^\mu \rangle_{S_n} = n!. \quad (3\text{-}6)$$

We note that there are $n - \varphi(n) - 1$ numbers k strictly between 1 and n so that $d_k \neq 1$, so substituting (3-4), (3-5), and (3-6) into (3-3) we have

$$\langle \phi, \chi^{(n,n)} \rangle_{S_{2n}} = \frac{1}{n n!} \big(n! + \varphi(n) n! + (n - \varphi(n) - 1) n! \big) = \frac{1}{n n!} n n! = 1. \quad \square$$

In the case of $n = 2$ it turns out that Propositions 3.1, 3.2, and 3.3 give a full decomposition. That is,

$$\mathrm{Ind}^{S_4}_{C_{S_4}((12))} 1_C = \chi^{(4)} + \chi^{(3,1)} + \chi^{(2,2)}.$$

We notice that our first three results all showed that there are certain irreducible characters appearing in the decomposition of ϕ that have constant or stable multiplicities, independent of n. Our next result shows that this is not the case for all constituents, but a closed-form formula for the multiplicity is known in some cases.

Proposition 3.4. *Let $n \geq 2$. Let $\chi^{(2n-2,2)}$ be the irreducible character of S_{2n} corresponding to $(2n-2, 2)$. Then*

$$\langle \phi, \chi^{(2n-2,2)} \rangle_{S_{2n}} = \begin{cases} \dfrac{n}{2} & \text{if } n \text{ is even,} \\[2mm] \dfrac{n-1}{2} & \text{if } n \text{ is odd.} \end{cases}$$

Proof. Throughout, $d_k = \gcd(n, k)$. Using Frobenius reciprocity we write

$$\langle \phi, \chi^{(2n-2,2)} \rangle_{S_{2n}} = \langle 1_C, \chi^{(2n-2,2)} \downarrow_C \rangle_C = \frac{1}{n n!} \sum_{k=0}^{n-1} \sum_{\pi \in S_n} \chi^{(2n-2,2)}((\sigma^k, \pi)).$$

If $n = 2$, we are done, by Proposition 3.3. Throughout the rest of the proof we assume $n \geq 3$. As in the proof of Proposition 3.3, we break the sum into three

pieces:

$$\langle \phi, \chi^{(2n-2,2)} \rangle_{S_{2n}}$$

$$= \frac{1}{nn!} \left(\sum_{\pi \in S_n} \chi^{(2n-2,2)}((1,\pi)) + \varphi(n) \sum_{\pi \in S_n} \chi^{(2n-2,2)}((\sigma,\pi)) \right.$$

$$\left. + \sum_{\substack{1<k<n \\ d_k \neq 1}} \sum_{\pi \in S_n} \chi^{(2n-2,2)}((\sigma^k,\pi)) \right)$$

$$= \frac{1}{nn!} \left(n! \langle \chi^{(n)}, \chi^{(2n-2,2)} \downarrow_{S_n} \rangle_{S_n} + \varphi(n) \sum_{\lambda \vdash n} \chi^{(2n-2,2)}_{(n,\lambda)} |K_\lambda| \right.$$

$$\left. + \sum_{\substack{1<k<n \\ d_k \neq 1}} \sum_{\lambda \vdash n} \chi^{(2n-2,2)}_{((\frac{n}{d_k})^{d_k},\lambda)} |K_\lambda| \right). \quad (3\text{-}7)$$

From Theorem 2.9, we have

$$\langle \chi^{(n)}, \chi^{(2n-2,2)} \downarrow_{S_n} \rangle_{S_n} = \binom{n}{2}. \quad (3\text{-}8)$$

Using Theorem 2.7 we write $\chi^{(2n-2,2)}_{(n,\lambda)} = \chi^{(n-2,2)}_\lambda$, so that

$$\sum_{\lambda \vdash n} \chi^{(2n-2,2)}_{(n,\lambda)} |K_\lambda| = \sum_{\lambda \vdash n} \chi^{(n-2,2)}_\lambda |K_\lambda| = n! \langle \chi^{(n)}, \chi^{(n-2,2)} \rangle_{S_n} = 0. \quad (3\text{-}9)$$

When n is even, $\frac{n}{2}$ divides n. Then $d_{\frac{n}{2}} = \frac{n}{2}$. We can then remove the 2-hook from bottom row of $(2n-2,2)$, and then successively remove $\frac{n}{2} - 1$ hooks of length 2 from the top row of $(2n-2,2)$. There are $\binom{n/2}{1} = \frac{n}{2}$ ways to do this. We combine this with Theorem 2.7 to see that

$$\sum_{\substack{1<k<n \\ d_k \neq 1}} \chi^{(2n-2,2)}_{((\frac{n}{d_k})^{d_k},\lambda)} = \frac{n}{2} \chi^{(n)} + \sum_{\substack{\mu \vdash n \\ \mu \neq (n)}} a_\mu \chi^\mu_\lambda$$

with $a_\mu \in \mathbb{Z}$. Then

$$\sum_{\substack{1<k<n \\ d_k \neq 1}} \sum_{\lambda \vdash n} \chi^{(2n-2,2)}_{((\frac{n}{d_k})^{d_k},\lambda)} |K_\lambda| = \sum_{\lambda \vdash n} \frac{n}{2} \chi^{(n)} |K_\lambda| + \sum_{\lambda \vdash n} \sum_{\substack{\mu \vdash n \\ \mu \neq (n)}} a_\mu \chi^\mu_\lambda |K_\lambda|$$

$$= \frac{n}{2} n! \langle \chi^{(n)}, \chi^{(n)} \rangle_{S_n} + \sum_{\substack{\mu \vdash n \\ \mu \neq (n)}} a_\mu \langle \chi^{(n)}, \chi^\mu \rangle_{S_n}$$

$$= \frac{n}{2} n!. \quad (3\text{-}10)$$

So in the case when n is even, substituting (3-8)–(3-10) into (3-7), we have

$$\langle \phi, \chi^{(2n-2,2)} \rangle_{S_{2n}} = \frac{1}{nn!}\left(\binom{n}{2}n! + \frac{n}{2}n!\right) = \frac{1}{n}\left(\binom{n}{2} + \frac{n}{2}\right) = \frac{1}{n}\left(\frac{n(n-1)}{2} + \frac{n}{2}\right)$$

$$= \frac{n-1}{2} + \frac{1}{2} = \frac{n}{2},$$

as desired. Now, when n is odd, 2 does not divide n. Then $\frac{n}{2}$ is not an integer and thus does not divide n. As a result, we cannot remove the hook of length 2 from the bottom row of $(2n-2,2)$. So when we apply Theorem 2.7, the trivial character does not appear in the decomposition and we have

$$\sum_{\substack{1<k<n \\ d_k \neq 1}} \chi^{(2n-2,2)}_{((\frac{n}{d_k})^{d_k},\lambda)} = \sum_{\substack{\mu \vdash n \\ \mu \neq (n)}} c_\mu \chi^\mu_\lambda \quad \text{with } c_\mu \in \mathbb{Z}.$$

Then

$$\sum_{\substack{1<k<n \\ d_k \neq 1}} \sum_{\lambda \vdash n} \chi^{(2n-2,2)}_{((\frac{n}{d_k})^{d_k},\lambda)} |K_\lambda| = \sum_{\lambda \vdash n} \sum_{\substack{\mu \vdash n \\ \mu \neq (n)}} c_\mu \chi^\mu_\lambda |K_\lambda|$$

$$= \sum_{\substack{\mu \vdash n \\ \mu \neq (n)}} n! \, c_\mu \langle \chi^{(n)}, \chi^\mu \rangle_{S_n} = 0. \qquad (3\text{-}11)$$

So then, substituting (3-8), (3-9), and (3-11) into (3-7), we have

$$\langle \phi, \chi^{(2n-2,2)} \rangle_{S_{2n}} = \frac{1}{nn!}\binom{n}{2}n! = \frac{1}{n}\binom{n}{2} = \frac{1}{n}\frac{n(n-1)}{2} = \frac{n-1}{2},$$

giving the result. $\qquad \square$

A theorem for the partitions $(2n-k,k)$. We now present a theorem that generalizes the previous propositions and gives a formula for the multiplicities of a number of the irreducible characters of S_{2n} appearing in the decomposition of ϕ.

Theorem 3.5. *Let $n \geq 2k$. Let $\chi^{(2n-k,k)}$ be the irreducible character of S_{2n} corresponding to $(2n-k,k)$. For $1 < h < n$, let $d_h = \gcd(n,h)$, and $l_k = kd_h/n$. Then*

$$\langle \phi, \chi^{(2n-k,k)} \rangle_{S_{2n}} = \frac{1}{n}\left(\binom{n}{k} + \sum_{\substack{1<h<n \\ d_h \neq 1 \\ \frac{n}{d_h} | k}} \binom{d_h}{l_k}\right).$$

Proof. With Frobenius reciprocity, we write

$$\langle \phi, \chi^{(2n-k,k)} \rangle_{S_{2n}} = \langle 1_C, \chi^{(2n-k,k)} \downarrow_C \rangle_C = \frac{1}{nn!} \sum_{j=0}^{n-1} \sum_{\pi \in S_n} \chi^{(2n-k,k)}((\sigma^j, \pi)).$$

As usual, we split the sum into three pieces:

$$\langle \phi, \chi^{(2n-k,k)} \downarrow_{S_{2n}} \rangle_{S_{2n}}$$

$$= \frac{1}{nn!} \left(\sum_{\pi \in S_n} \chi^{(2n-k,k)}((1,\pi)) \right.$$

$$+ \varphi(n) \sum_{\pi \in S_n} \chi^{(2n-k,k)}((\sigma,\pi)) + \sum_{\substack{1 < h < n \\ d_h \neq 1}} \sum_{\pi \in S_n} \chi^{(2n-k,k)}((\sigma^h,\pi)) \left. \right)$$

$$= \frac{1}{nn!} \left(n! \langle \chi^{(n)}, \chi^{(2n-k,k)} \downarrow_{S_n} \rangle_{S_n} \right.$$

$$+ \varphi(n) \sum_{\lambda \vdash n} \chi^{(2n-k,k)}_{(n,\lambda)} |K_\lambda| + \sum_{\substack{1 < h < n \\ d_h \neq 1}} \sum_{\lambda \vdash n} \chi^{(2n-k,k)}_{((\frac{n}{d_h})^{d_h},\lambda)} |K_\lambda| \left. \right). \qquad (3\text{-}12)$$

Since $n \geq 2k$, we can remove the k blocks from the bottom row of $(2n-k,k)$ and remove $n-k$ blocks from the top row of $(2n-k,k)$, which leaves n blocks remaining. We can do this removal in $\binom{n}{k}$ ways, so, with 2.9, we have

$$\langle \chi^{(n)}, \chi^{(2n-k,k)} \downarrow_{S_n} \rangle_{S_n} = \binom{n}{k}. \qquad (3\text{-}13)$$

Theorem 2.7 gives

$$\chi^{(2n-k,k)}_{(n,\lambda)} = \chi^{(n-k,k)}_{\lambda},$$

since $n \geq 2k$. Then

$$\sum_{\lambda \vdash n} \chi^{(2n-k,k)}_{(n,\lambda)} |K_\lambda| = \sum_{\lambda \vdash n} \chi^{(n-k,k)}_{\lambda} |K_\lambda| = n! \langle \chi^{(n)}, \chi^{(n-k,k)} \rangle_{S_n} = 0. \qquad (3\text{-}14)$$

Now suppose there is some h so that $d_h \neq 1$. Then σ^h is a product of d_h $\frac{n}{d_h}$-cycles. If π is of cycle type λ then

$$\chi^{(2n-k,k)}((\sigma^h,\pi)) = \chi^{(2n-k,k)}_{((\frac{n}{d_h})^{d_h},\lambda)}. \qquad (3\text{-}15)$$

By Theorem 2.7, in order for $\chi^{(n)}$ to have nonzero multiplicity in the decomposition of the right-hand side of (3-15), we have to be able to remove the k-hook from the bottom row of $(2n-k,k)$. So if $\frac{n}{d_h}$ does not divide k then this is not possible. Then in this case

$$\chi^{(2n-k,k)}_{((\frac{n}{d_h})^{d_h},\lambda)} = \sum_{\substack{\mu \vdash n \\ \mu \neq (n)}} a_\mu \chi^{\mu}_{\lambda} \qquad (3\text{-}16)$$

with $a_\mu \in \mathbb{Z}$. Now, suppose that, for some h, $d_h \neq 1$, and furthermore that $\frac{n}{d_h}$ divides k. Then we can successively remove the l_k hooks of length $\frac{n}{d_h}$ from

the bottom row of $(2n - k, k)$ and remove the $d_h - l_k$ hooks of length $\frac{n}{d_h}$ from the top row of $(2n - k, k)$, which will result in $\chi^{(n)}$ having positive multiplicity in the aforementioned decomposition. In fact, a simple counting argument via Theorem 2.9 shows the exact multiplicity will be $\binom{d_h}{l_k}$. Then in this case

$$\chi^{(2n-k,k)}_{((\frac{n}{d_h})^{d_h},\lambda)} = \binom{d_h}{l_k}\chi^{(n)}_\lambda + \sum_{\substack{\mu \vdash n \\ \mu \neq (n)}} c_\mu \chi^\mu_\lambda \tag{3-17}$$

with $c_\mu \in \mathbb{Z}$. Then (3-16) and (3-17) give

$$\sum_{\substack{1 < h < n \\ d_h \neq 1}} \sum_{\lambda \vdash n} \chi^{(2n-k,k)}_{((\frac{n}{d_h})^{d_h},\lambda)} |K_\lambda|$$

$$= \sum_{\substack{1 < h < n \\ d_h \neq 1 \\ \frac{n}{d_h} \nmid k}} \sum_{\lambda \vdash n} \sum_{\substack{\mu \vdash n \\ \mu \neq (n)}} a_\mu \chi^\mu_\lambda |K_\lambda| + \sum_{\substack{1 < h < n \\ d_h \neq 1 \\ \frac{n}{d_h} \mid gk}} \sum_{\lambda \vdash n} \binom{d_h}{l_k}\chi^{(n)}_\lambda |K_\lambda|$$

$$+ \sum_{\substack{1 < h < n \\ d_h \neq 1 \\ \frac{n}{d_h} \mid gk}} \sum_{\lambda \vdash n} \sum_{\substack{\mu \vdash n \\ \mu \neq (n)}} c_\mu \chi^\mu_\lambda |K_\lambda|$$

$$= \sum_{\substack{1 < h < n \\ d_h \neq 1 \\ \frac{n}{d_h} \mid gk}} \sum_{\lambda \vdash n} \binom{d_h}{l_k}\chi^{(n)}_\lambda |K_\lambda| = \sum_{\substack{1 < h < n \\ d_h \neq 1 \\ \frac{n}{d_h} \mid gk}} \binom{d_h}{l_k}n!. \tag{3-18}$$

Substituting (3-13), (3-14), (3-18) into (3-12) we have

$$\langle \phi, \chi^{(2n-k,k)} \rangle_{S_{2n}} = \frac{1}{nn!}\left(\binom{n}{k}n! + \sum_{\substack{1 < h < n \\ d_h \neq 1 \\ \frac{n}{d_h} \mid k}} \binom{d_h}{l_k}n!\right)$$

$$= \frac{1}{n}\left(\binom{n}{k} + \sum_{\substack{1 < h < n \\ d_h \neq 1 \\ \frac{n}{d_h} \mid k}} \binom{d_h}{l_k}\right), \tag{3-19}$$

as claimed. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

4. Future problems

The preceding work is only the beginning of a large selection of problems to be worked out. It is possible that there are more stable multiplicities (independent of n) in this decomposition. Also, the multiplicities and formulas found here only

cover a small number of partitions and therefore characters. One may find that *all* characters have a stable or closed-form formula for their multiplicities. Note that in this paper we only discuss the trivial character of C, and much can be learned from studying the decomposition of the nontrivial characters of C when induced up to S_{2n}, which arise in braid group cohomology. It may be possible to learn more by first decomposing the character $\text{Ind}_C^{S_n} \psi$, studying this character, and then inducing the resulting constituents up to S_{2n}.

Acknowledgements

References

[Church and Farb 2010] T. Church and B. Farb, "Representation theory and homological stability", preprint, 2010. arXiv 1008.1368

[Dummit and Foote 2004] D. S. Dummit and R. M. Foote, *Abstract algebra*, 3rd ed., John Wiley & Sons, Hoboken, NJ, 2004. MR 2007h:00003 Zbl 1037.00003

[Hemmer 2011] D. J. Hemmer, "Stable decompositions for some symmetric group characters arising in braid group cohomology", *J. Combin. Theory Ser. A* **118**:3 (2011), 1136–1139. MR 2012a:20021 Zbl 1231.20011

[James and Liebeck 2001] G. James and M. Liebeck, *Representations and characters of groups*, 2nd ed., Cambridge University Press, 2001. MR 2002h:20010 Zbl 0981.20004

[Jöllenbeck and Schocker 2000] A. Jöllenbeck and M. Schocker, "Cyclic characters of symmetric groups", *J. Algebraic Combin.* **12**:2 (2000), 155–161. MR 2001k:05207 Zbl 0979.20017

[Kraśkiewicz and Weyman 2001] W. Kraśkiewicz and J. Weyman, "Algebra of coinvariants and the action of a Coxeter element", *Bayreuth. Math. Schr.* 63 (2001), 265–284. MR 2002j:20026 Zbl 1037.20012

[Sagan 2001] B. E. Sagan, *The symmetric group: Representations, combinatorial algorithms, and symmetric functions*, 2nd ed., Graduate Texts in Mathematics **203**, Springer, New York, 2001. MR 2001m:05261 Zbl 0964.05070

jjricci@buffalo.edu   *Mathematics Department, University at Buffalo, SUNY,
                                        Buffalo, NY 14260, United States*

msp

On the geometric deformations of functions in $L^2[D]$

Luis Contreras, Derek DeSantis and Kathryn Leonard

(Communicated by David Royal Larson)

We derive a formal relationship between the coefficients of a function expanded in either the Legendre basis or Haar wavelet basis, before and after a polynomial deformation of the function's domain. We compute the relationship of coefficients explicitly in three cases: linear deformation with Haar basis, linear deformation with Legendre basis, and polynomial deformation with Legendre basis.

1. Introduction

This paper explores the relationship between Schauder coefficients of a function before and after the domain of that function has been deformed in some reasonably well-behaved manner. As an analogy, one may think of a function as a melody recorded on an LP, and its domain as the position in the groove on the LP. The groove will become deformed if the LP is left in the sun, but the melody played on the LP after deformation will be related to the original melody. We are interested in understanding that relationship. Our results are a preliminary step toward addressing the inverse question of how to recover information about the undeformed function given the deformed function and an unknown deformation.

More formally, let $\mathcal{W} = \{w : D \to D \mid w$ is a diffeomorphism$\}$ be a class of diffeomorphisms defined on a closed subinterval $D \subset \mathbb{R}$. Then each $w \in \mathcal{W}$ defines a function F_w on $L^2[D]$, where $F_w(f) = f \circ w$. Below, we provide necessary background information to pose our question in terms of coefficients of elements in $L^2[D]$. In Section 2, we derive a general relationship between the coefficients of f, w, and $g = F_w(f)$. In Section 3, we compute precise relationships between coefficients of f and g in the Legendre and Haar wavelet bases for linear deformations, and in Section 4, in the Legendre basis for polynomial deformations.

1.1. *Background.* For the Hilbert space $L^2[D] = \{f : D \to \mathbb{R} \mid \int_D f^2 < \infty\}$, recall that the inner product is given by $\langle f, g \rangle = \int_D fg$. Therefore, given an orthonormal basis $\{\phi_i(x)\}_{i=0}^\infty$ for $L^2[D]$, the Schauder coefficients $\{a_i\}$ corresponding to a function expanded in that basis, $f(x) = \sum_{i=0}^\infty a_i \phi_i(x)$, can be computed by $a_i = \int_D f(x)\phi_i(x)\,dx$ [Kreyszig 1989].

We will be exploring two orthonormal bases in our work: the Legendre basis, which is a basis of polynomials, and the Haar wavelet basis, a basis that localizes in scale and location. As noted above, domain deformation corresponds to composition of functions. The Legendre basis has the advantage that computations involving composition with polynomial deformations are straightforward. On the other hand, because the support of each basis function is the entire domain D, localized deformations will produce changes in every Legendre coefficient. The Haar wavelet basis has the opposite problem: local deformations will change only the subset of coefficients corresponding to that locale, but composing basis functions with polynomial deformations is computationally intimidating. Examined together, however, these two bases provide a wide view of possible behaviors. We now define each basis formally.

1.1.1. *Legendre basis for $L^2[-1, 1]$.* The Legendre basis arises by applying the Gram–Schmidt orthonormalization process to the simplest basis for $L^2[-1, 1]$, the monomials $\{x^i\}_{i=0}^\infty$. For $D = [-1, 1]$, the resulting basis is as below (though choosing a different D will produce a different normalizing constant K):

$$\psi_i(x) = \begin{cases} \sqrt{\dfrac{2i+1}{2}} \displaystyle\sum_{n=0}^N (-1)^n \dfrac{(2i-2n)!}{2^i n!(i-n)!(i-2n)!} x^{i-2n} & \text{for } -1 \le x \le 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $N = i/2$ when i is even, and $N = (i-1)/2$ when i is odd [Jackson 2004]. Rewriting the normalizing constant

$$K_{in} = \sqrt{\frac{2i+1}{2}} (-1)^n \frac{(2i-2n)!}{2^i n!(i-n)!(i-2n)!},$$

our basis becomes

$$\psi_i(x) = \sum_{n=0}^N K_{in} x^{i-2n}. \tag{1}$$

A function $f(x) \in L^2[-1, 1]$ can therefore be written as

$$f(x) = \sum_{i=0}^\infty a_i \sum_{n=0}^N K_{in} x^{i-2n} = \sum_i \sum_n a_i K_{in} x^{i-2n}.$$

1.1.2. *Haar basis for $L^2[0, 1]$.* The Haar wavelet basis is generated by shifting and scaling the simplest mother wavelet,

$$\psi(x) = \begin{cases} 1 & \text{for } 0 \leq x < \frac{1}{2}, \\ -1 & \text{for } \frac{1}{2} \leq x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

which can be thought of as a coarse piecewise constant approximation to a sine curve. After scaling and shifting, the resulting orthonormal basis is given by

$$\psi_{ij}(x) = \begin{cases} 2^{i/2} & \text{for } \frac{j}{2^i} \leq x < \frac{j+1/2}{2^i}, \\ -2^{i/2} & \text{for } \frac{j+1/2}{2^i} \leq x < \frac{j+1}{2^i}, \\ 0 & \text{otherwise,} \end{cases}$$

where $i \in \mathbb{N}$ and $0 \leq j \leq 2^i - 1$ [Radunović 2009].

2. General relationships of coefficients

Our first result presents a general relationship between Schauder coefficients of f and those of g.

Theorem 1. *Consider $f(x) \in L^2[D]$, where $D \subset \mathbb{R}$ is a a closed interval, and let $w(x) = h^{-1}(x) : D \to D$ be a diffeomorphism. Set $g(x) = f \circ w(x)$. Then for $f(x) = \sum_i a_i \psi_i(x)$, where $\{\psi_i\}$ is an orthonormal basis for $L^2[D]$,*

$$g(x) = \sum_i c_i \psi_i(x) = \sum_i \sum_j \alpha_{ij} a_j \psi_i(x),$$

where $\alpha_{ij} = \langle \psi_j \circ w(x), \psi_i(x) \rangle_{L^2}$.

Proof. We claim that $g \in L^2(D)$. Because w is a diffeomorphism, w' is continuous and nonvanishing on D. Therefore, $1/w'$ is also continuous on D and thus bounded above by some $M < \infty$. We then have $\int_D g^2 = \int_D (f \circ w)^2 = \int f^2/w' \leq M \|f\|_2^2 < \infty$, and so $g \in L^2[D]$.

Thus, we can write $g(x)$ as the convergent series $\sum_i c_i \psi_i(x)$, where $c_i = \langle g, \psi_i \rangle$. Remembering that $g = f \circ w = \sum_j a_j(\psi_j \circ w)$, we have

$$c_i = \langle g, \psi_i \rangle = \langle f \circ w, \psi_i \rangle$$
$$= \left\langle \sum_j a_j(\psi_j \circ w), \psi_i \right\rangle = \sum_j a_j \langle \psi_j \circ w, \psi_i \rangle = \sum_j a_j \alpha_{ij}. \qquad \square$$

Note that the coefficients $\{\alpha_{ij}\}$ can be computed independently of f. Given a deformation w, these may be computed and reused for multiple choices of f. Alas, such a clean theorem requires dues to be paid elsewhere. We will see below the challenges of computing the $\{\alpha_{ij}\}$ coefficients in specific cases.

3. Explicit relationships: linear deformations

3.1. *Linear deformations and the Legendre basis for $L^2[-1, 1]$.* We first examine deformations of the form $w(x) = \beta x$, with $0 < \beta < 1$, for $D = [-1, 1]$. We are cheating slightly here, as $h = w^{-1}$ maps D to a larger interval $D \subset h(D)$, and so the setting of this first example does not match with Theorem 1. Nonetheless, the results for linear w will be helpful in understanding the results for polynomial w in Section 4, and so we persevere. We start with a simple fact from calculus:

Fact. *For $A = [-a, a]$ and t odd, $\int_A x^t \, dx = 0$.* $\qquad\qquad\square$

Theorem 2. *Following Theorem 1, we take $D = [-1, 1]$, $\{\psi_i(x)\}$ as the Legendre basis, and $w(x) = \beta x$, $\beta > 0$. Then*

$$
\alpha_{ij} =
\begin{cases}
2 \sum_{n,m=0}^{N,M} \dfrac{K_{in} K_{jm} \beta^{i-2n}}{(i - 2n) + (j - 2m) + 1} & \text{if } i + j \text{ is even} \\[2mm]
0 & \text{otherwise.}
\end{cases}
$$

Proof. Expanding a function f in the Legendre basis, we can write $f(x)$ as $\sum_i a_i \sum_{n=0}^N K_{in} x^{i-2n}$, where $N = i/2$ when i is even and $N = (i-1)/2$ when i is odd. We are concerned with $g(x) = f(w(x)) = \sum_i a_i \psi_i(w(x))$, where

$$
\psi_i(w(x)) = \psi_i(\beta x) = \sum_{n=0}^N K_{in} (\beta x)^{i-2n} = \sum_{n=0}^N K_{in} x^{i-2n} \beta^{i-2n}.
$$

Therefore,

$$
g(x) = \sum_i a_i \sum_{n=0}^N K_{in} x^{i-2n} \beta^{i-2n}.
$$

Substituting in βx, we obtain the following formula for $\{\alpha_{ij}\}$:

$$
\begin{aligned}
\alpha_{ij} &= \langle \psi_i(\beta x), \psi_j(x) \rangle \\
&= \int_{-1}^1 \left(\sum_{n=0}^N K_{in} x^{i-2n} \beta^{i-2n} \right) \left(\sum_{m=0}^M K_{jm} x^{j-2m} \right) dx \\
&= \int_{-1}^1 \sum_{n,m=0}^{N,M} (K_{in} x^{i-2n} \beta^{i-2n})(K_{jm} x^{j-2m}) \, dx \\
&= \sum_{n,m=0}^{N,M} \int_{-1}^1 (K_{in} x^{i-2n} \beta^{i-2n})(K_{jm} x^{j-2m}) \, dx. \\
&= \sum_{n,m=0}^{N,M} \int_{-1}^1 K_{in} K_{jm} x^{i-2n+j-2m} \beta^{i-2n} \, dx
\end{aligned}
$$

In view of the Fact quoted above, if $i + j$ is odd, the integral is zero. Otherwise,

$$\alpha_{ij} = \sum_{n,m=0}^{N,M} \int_{-1}^{1} K_{in} K_{jm} x^{i-2n+j-2m} \beta^{i-2n} \, dx = 2 \sum_{n,m=0}^{N,M} \frac{K_{in} K_{jm} \beta^{i-2n}}{(i-2n)+(j-2m)+1}. \qquad \square$$

3.2. *Linear deformations and the Haar basis for $L^2[0, 1]$.* We again examine linear deformations of the form $w(x) = \beta x$, now with $\beta > 0$ and $D = [0, 1]$. Note that for the Haar wavelet basis, each basis element has two indices: one for scale and one for location. Hence, the $\{\alpha_{ij}\}$ coefficients defined in Theorem 1 become $\{\alpha_{ijkl}\} = \langle \psi_{ij} \circ w, \psi_{kl} \rangle$.

As before, we must compute

$$\psi_{ij}(w(x)) = \psi_{ij}(\beta x) = \begin{cases} 2^{i/2} & \text{for } \frac{j}{2^i} \le \beta x < \frac{j+1/2}{2^i} \\ -2^{i/2} & \text{for } \frac{j+1/2}{2^i} \le \beta x < \frac{j+1}{2^i} \\ 0 & \text{otherwise.} \end{cases}$$

$$= \begin{cases} 2^{i/2} & \text{for } \frac{j}{\beta 2^i} \le x < \frac{j+1/2}{\beta 2^i} \\ -2^{i/2} & \text{for } \frac{j+1/2}{\beta 2^i} \le x < \frac{j+1}{\beta 2^i} \\ 0 & \text{otherwise.} \end{cases}$$

Let

$$I_{ij}^+ = \left[\frac{j}{\beta 2^i}, \frac{j+1/2}{\beta 2^i} \right) \quad \text{and} \quad I_{ij}^- = \left[\frac{j+1/2}{\beta 2^i}, \frac{j+1}{\beta 2^i} \right),$$

the regions where $\psi_{ij}(w(x)) > 0$ and $\psi_{ij}(w(x)) < 0$, respectively. Similarly, let

$$I_{kl}^+ = \left[\frac{l}{2^k}, \frac{l+1/2}{2^k} \right) \quad \text{and} \quad I_{kl}^- = \left[\frac{l+1/2}{2^k}, \frac{l+1}{2^k} \right).$$

Note that a particular α_{ijkl} will be nonzero only if (a) $(I_{ij}^+ \cup I_{ij}^-) \cap I_{kl}^+ \ne \varnothing$ and $(I_{ij}^+ \cup I_{ij}^-) \cap I_{kl}^- \ne \varnothing$ and (vice versa) (b) $(I_{kl}^+ \cup I_{kl}^-) \cap I_{ij}^+ \ne \varnothing$ and $(I_{kl}^+ \cup I_{kl}^-) \cap I_{ij}^- \ne \varnothing$. Otherwise, α_{ijkl} will vanish; either the supports will be disjoint, or the support of one will be contained entirely in the positive or negative domain of the other. Analyzing the possibilities for nonzero values of α_{ijkl} produces the following theorem.

Theorem 3. *Following Theorem 1, we take $D = [0, 1]$, $\{\psi_{ij}(x)\}$ as the Haar wavelet basis, and $w(x) = \beta x$, $\beta > 0$. Then nonzero values for α_{ijkl} are of the form*

$$\alpha_{ijkl} = \sum_{m=1}^{3} \xi_m 2^{m-k-2} + \tilde{\xi}_m 2^{m-i-2},$$

where $\xi_1 = 1 \pm \beta$, $\xi_2 \in \{-\beta, \pm l\beta, (1+l)\beta, \pm(3l+1)\beta\}$, $\xi_3 \in \{\pm l\beta, (1+l)\beta\}$, $\tilde{\xi}_1 = 0$, $\tilde{\xi}_2 \in \{\pm 1, \pm j, \pm(1+j), 3j, -(3j+1)\}$, and $\tilde{\xi}_3 \in \{1, \pm j, (1+j)\}$.

Proof. Expanding some f in the Haar basis, we can write $f(x) = \sum\limits_{i=0}^{\infty} \sum\limits_{j=0}^{2^i-1} a_{ij} \psi_{ij}(x)$. Therefore,

$$g(x) = f(w(x)) = \sum_{i=0}^{\infty} \sum_{j=0}^{2^i-1} a_{ij} \psi_{ij}(w(x)) = \sum_{i=0}^{\infty} \sum_{j=0}^{2^i-1} a_{ij} \psi_{ij}(\beta x).$$

The formula for α_{ijkl} is then

$$\alpha_{ijkl} = \langle \psi_{ij}(\beta x), \psi_{kl}(x) \rangle$$
$$= \int_{(I_{ij}^+ \cap I_{kl}^+) \cup (I_{ij}^- \cap I_{kl}^-)} 2^{i/2} 2^{k/2}\, dx - \int_{(I_{ij}^+ \cap I_{kl}^-) \cup (I_{ij}^- \cap I_{kl}^+)} 2^{i/2} 2^{k/2}\, dx$$
$$= 2^{(i+k)/2} \left(\int_{(I_{ij}^+ \cap I_{kl}^+) \cup (I_{ij}^- \cap I_{kl}^-)} dx - \int_{(I_{ij}^+ \cap I_{kl}^-) \cup (I_{ij}^- \cap I_{kl}^+)} dx \right)$$
$$= 2^{(i+k)/2} \left[\mu\big((I_{ij}^+ \cap I_{kl}^+) \cup (I_{ij}^- \cap I_{kl}^-)\big) - \mu\big((I_{ij}^+ \cap I_{kl}^-) \cup (I_{ij}^- \cap I_{kl}^+)\big) \right],$$

where μ is the standard Lebesgue measure. Hence, to compute α_{ijkl}, we must compute $M = \mu\big((I_{ij}^+ \cap I_{kl}^+) \cup (I_{ij}^- \cap I_{kl}^-)\big) - \mu\big((I_{ij}^+ \cap I_{kl}^-) \cup (I_{ij}^- \cap I_{kl}^+)\big)$. From the 14 possible arrangements of the values $\{\frac{l}{2^k}, \frac{l+1/2}{2^k}, \frac{l+1}{2^k}, \frac{j}{\beta 2^i}, \frac{j+1/2}{\beta 2^i}, \frac{j+1}{\beta 2^i}\}$ satisfying (a) and (b) as in the discussion preceding Theorem 3, we find possible values for M as follows. Given positive integers $\{ijkl\}$ corresponding to α_{ijkl}, the value of $\beta 2^{i+k+2} M \neq 0$ is one of

- $(1+j)2^{k+3} - l\beta 2^{i+3} - \beta 2^{i+2}$
- $(1+j)2^{k+2} - l\beta 2^{i+2}$
- $-j2^{k+3} + l\beta 2^{i+3} + (1 + \beta 2^{i+1}\}$
- $-(3j+1)2^{k+2} + (3l+1)\beta 2^{i+2} + (1+\beta)2^{i+1}$
- $-j2^{k+2} + l\beta 2^{i+2} + (1-\beta)2^{i+1}$
- $\pm j2^{k+2} \mp (1+l)\beta 2^{i+2}$
- $-(1+j)2^{k+2} + (1+l)\beta 2^{i+2}$
- $-j2^{k+3} - 2^{k+2} + (1+l)\beta 2^{i+3}$
- $j2^{k+3} + 2^{k+2} - l\beta 2^{i+3}$
- $2^{k+3} + 3j2^{k+2} - (3l+1)\beta 2^{i+2}$
- $j2^{k+2} - l\beta 2^{i+2}$
- $-(1+j)2^{k+2} + l\beta 2^{i+2}$
- $(1+j)2^{k+2} - (1+l)\beta 2^{i+2}$.

Substituting these values for M into the formula for α_{ijkl} gives the desired result. $\square$

4. Explicit relationships: polynomial deformations and the Legendre basis

Because the Legendre basis is a basis of polynomials, it is less challenging to compute values for $\{\alpha_{ij}\}$ when w is a polynomial than it would be for a nonpolynomial basis such as the Haar basis. We now consider deformations $w(x) = \sum_{s=0}^{v} \beta_s x^s$, where the $\{\beta_s\}$ are chosen so that $w(x)$ maps $[-1, 1]$ onto itself diffeomorphically and $dw/dx > 0$. This increase in complexity of the deformations requires careful accounting, as we shall see below.

As before, we compute

$$\psi_i(w(x)) = \psi_i\left(\sum_{s=0}^{v} \beta_s x^s\right) = \sum_{n=0}^{N} K_{in}\left(\sum_{s=0}^{v} \beta_s x^s\right)^{i-2n}$$

$$= \sum_{n=0}^{N} K_{in}\left(\sum_{p_0+p_1+\cdots+p_v=i-2n} \binom{i-2n}{p_0, p_1, \ldots, p_v}(\beta_0 x^0)^{p_0}(\beta_1 x^1)^{p_1}\cdots(\beta_v x^v)^{p_v}\right)$$

$$= \sum_{n=0}^{N} K_{in}\left(\sum_{p_0+p_1+\cdots+p_v=i-2n} \binom{i-2n}{p_0, p_1, \ldots, p_v}\left(\prod_{s=0}^{v}\beta_s^{p_s}\right)x^{\sum_{s=0}^{v} sp_s}\right)$$

$$= \sum_{n=0}^{N}\sum_{P} K_{in}\binom{i-2n}{p_0, p_1, \ldots, p_v}\left(\prod_{s=0}^{v}\beta_s^{p_s}\right)x^{\sum_{s=0}^{v} sp_s}$$

using the multinomial theorem, where $P = p_0 + p_1 + \cdots + p_v$ is the collective sum of partitions of $i - 2n$. Therefore,

$$g(x) = f(w(x)) = \sum_{i} a_i \sum_{n=0}^{N}\sum_{P} K_{in}\binom{i-2n}{p_0, p_1, \ldots, p_v}\left(\prod_{s=0}^{v}\beta_s^{p_s}\right)x^{\sum_{s=0}^{v} sp_s}.$$

In order to apply the Fact to compute $\langle\psi_j(w(x)), \psi_i(x)\rangle$, we must identify which of the powers of x, given by $\sum_{s=0}^{v} sp_s$, are even and which are odd. Certainly, when s is even, sp_s will be even. We rewrite

$$\sum_{s=0}^{v} sp_s = \sum_{t=0}^{\lfloor v/2 \rfloor}\left(2tp_{2t} + (2t+1)p_{2t+1}\right) = \sum_{t=0}^{\lfloor v/2 \rfloor} 2tp_{2t} + \sum_{t=0}^{\lfloor v/2 \rfloor}(2t+1)p_{2t+1}.$$

Analyzing the sum over odd $s = 2t+1$, we see that if p_{2t+1} is even for a given t, the product $(2t+1)p_{2t+1}$ will be even. In other words, the parity of the total exponent $\sum_{s=0}^{v} sp_s$ is determined entirely by the parity of the number of odd-indexed elements of the partition that are themselves odd. More precisely, let N_P be the number of odd-valued elements in the set $\{p_{2t+1}\}$. If N_P is odd, then $\sum_{t=0}^{\lfloor v/2 \rfloor}(2t+1)p_{2t+1}$ will sum an odd number of odd elements, and will therefore be odd. If N_P is even, $\sum_{t=0}^{\lfloor v/2 \rfloor}(2t+1)p_{2t+1}$ will sum an even number of odd elements, and will therefore be even. We have proved the following lemma.

Lemma. *Let $P : p_1 + \cdots + p_v = i - 2n$ be a particular choice of partition. Then the value of $\sum_{s=0}^{v} sp_s$ will be even if N_P, the number of odd-indexed, odd-valued elements of P, is even, or odd if N_P is odd.*

We now state the result for polynomial deformations.

Theorem 4. *Following Theorem 1, we take $D = [-1, 1]$, $\{\psi_i(x)\}$ as the Legendre basis, and $w(x) = \sum_{s=0}^{v} \beta_s x^s$ to be monotone increasing on D. Then*

$$\alpha_{ij} = 2 \sum_{n,m=0}^{N,M} \sum_{P,2|j+N_P} \binom{i - 2n}{p_0,\, p_1,\, \ldots,\, p_v} \frac{K_{in} K_{jm} \left(\prod_{s=0}^{v} \beta_s^{\,p_s} \right)}{j - 2m + 1 + \sum_{s=0}^{v} sp_s}.$$

Proof. Calculating α_{ij}, we find

$$\alpha_{ij} = \langle \psi_i(w(x)), \psi_j(x) \rangle$$

$$= \int_{-1}^{1} \left[\sum_{n=0}^{N} \sum_{P}^{i-2n} K_{in} \binom{i - 2n}{p_0,\, p_1,\, \ldots,\, p_v} \left(\prod_{s=0}^{v} \beta_s^{\,p_s} \right) x^{\sum_{s=0}^{v} sp_s} \right] \left[\sum_{m=0}^{M} K_{jm} x^{j-2m} \right] dx$$

$$= \sum_{n,m=0}^{N,M} \sum_{P}^{i-2n} K_{in} K_{jm} \binom{i - 2n}{p_0,\, p_1,\, \ldots,\, p_v} \left(\prod_{s=0}^{v} \beta_s^{\,p_s} \right) \int_{-1}^{1} x^{j-2m+\sum_{s=0}^{v} sp_s}\, dx.$$

Each integral term of the sum will vanish or not depending on the parity of $j - 2m + \sum_{s=0}^{v} sp_s$. Because $2m$ is always even, we focus on the parity of $j + \sum_{s=0}^{v} sp_s$. For each α_{ij}, j is fixed along with its parity. From the discussion leading up to Theorem 4, we know that N_P determines the parity of $\sum_{s=0}^{v} sp_s$. Putting this together, we see that the exponent $j - 2m + \sum_{s=0}^{v} sp_s$ will be odd (and so will have vanishing integral) when $j + N_P$ is odd. When $j + N_P$ is even, however, the exponent will be even and the integral nonzero. $\square$

5. Conclusion and future work

Based on the computational challenges apparent in the few simple examples given in this paper, we believe there are very few cases where the coefficients $\{\alpha_{ij}\}$ that capture the relationship between the deformed and undeformed function can be computed explicitly. Nonetheless, we would like to be able to say something in other situations. Currently, we are exploring distributions of coefficients of periodic functions after deformation by randomly generated b-splines with between 5 and 25 knots. We hope to make conjectures based on those empirical results about what we can realistically say mathematically. Because of the highly structured nature of periodic functions, we expect meaningful results. For example, since the oscillations of a periodic function cannot change in number or amplitude after composition with a deformation, there should be a formulation for a wavelet basis that relates scale and location of periodic behavior with the local energy of a deformation.

The motivation for this project comes from a similar problem in two dimensions related to modeling textures in images [Liu et al. 2004a; 2004b; Park et al. 2009]. When a periodic texture such as a wallpaper pattern appears in an image, it is often not periodic within the image. That is, geometric distortions arising from lighting, occlusion, or projection of a three-dimensional object onto the two-dimensional image plane, create a near-periodic texture in the image. To recognize the periodic structures in these distorted textures requires solving this problem: given a deformed near-periodic function, what is the underlying periodic function and the associated deformation? This inverse problem is ill-posed, but our work gives insight into a similar problem in one dimension. Future work will focus on examining that inverse problem in the one-dimensional setting and deriving similar results to the ones in this paper for functions on $\mathbb{R}^2$.

References

[Jackson 2004] D. Jackson, *Fourier series and orthogonal polynomials*, Dover, Mineola, NY, 2004. MR 2005g:42001 Zbl 1084.42001

[Kreyszig 1989] E. Kreyszig, *Introductory functional analysis with applications*, Wiley, New York, 1989. MR 90m:46003 Zbl 0706.46001

[Liu et al. 2004a] Y. Liu, R. T. Collins, and Y. Tsin, "A computational model for periodic pattern perception based on frieze and wallpaper groups", *IEEE Trans. Pattern Anal. Mach. Intell.* **26**:3 (2004), 354–371.

[Liu et al. 2004b] Y. Liu, W.-C. Lin, and J. Hays, "Near-regular texture analysis and manipulation", *ACM Trans. Graph.* **23**:3 (2004), 368–376.

[Park et al. 2009] M. Park, K. Brocklehurst, R. T. Collins, and Y. Liu, "Deformed lattice detection in real-world images using mean-shift belief propagation", *IEEE Trans. Pattern Anal. Mach. Intell.* **31**:10 (2009), 1804–1816.

[Radunović 2009] D. P. Radunović, *Wavelets: from math to practice*, Springer, Berlin, 2009. MR 2011h:42001 Zbl 1168.94300

luisdavidcz@aol.com   *Mathematics and Applied Physics, California State University, Channel Islands, Camarillo, CA 93012, United States*

derek.desantis23@gmail.com   *Mathematics and Applied Physics, University of Nebraska, Lincoln, Lincoln, NE 68521, United States*

kleonard.ci@gmail.com   *Department of Mathematics, California State University, Channel Islands, 1 University Dr, Camarillo, CA 93012, United States*

Spectral characterization for von Neumann's iterative algorithm in $\mathbb{R}^n$

Rudy Joly, Marco López, Douglas Mupasiri and Michael Newsome

(Communicated by Jim Haglund)

Our work is motivated by a theorem proved by von Neumann: Let S_1 and S_2 be subspaces of a closed Hilbert space X and let $x \in X$. Then

$$\lim_{k \to \infty} (P_{S_2} P_{S_1})^k(x) = P_{S_1 \cap S_2}(x),$$

where P_S denotes the orthogonal projection of x onto the subspace S. We look at the linear algebra realization of the von Neumann theorem in $\mathbb{R}^n$. The matrix A that represents the composition $P_{S_2} P_{S_1}$ has a form simple enough that the calculation of $\lim_{k \to \infty} A^k x$ becomes easy. However, a more interesting result lies in the analysis of eigenvalues and eigenvectors of A and their geometrical interpretation. A characterization of such eigenvalues and eigenvectors is shown for subspaces with dimension $n - 1$.

1. Introduction

In Euclidean n-space, we wish to find the point x_∞ in the intersection of two $(n-1)$-dimensional subspaces, S_1 and S_2, that is closest to an initial point x_0 in $\mathbb{R}^n$. That is, we want $x_\infty \in S_1 \cap S_2$ to be such that

$$\|x_0 - x_\infty\| \leq \|x_0 - y\| \quad \text{for all } y \in S_1 \cap S_2.$$

We call x_∞ the orthogonal projection of x_0 onto $S_1 \cap S_2$. We start by stating von Neumann's theorem; see [Deutsch 2001], for example.

Theorem 1. *Let S_1 and S_2 be subspaces of a closed Hilbert space X and let $x \in X$. Then*

$$\lim_{k \to \infty} (P_{S_2} P_{S_1})^k(x) = P_{S_1 \cap S_2}(x), \tag{1-1}$$

where P_S denotes the orthogonal projection onto the subspace S.

Von Neumann's theorem provides an iterative procedure (left-hand side of (1-1)) to find the orthogonal projection of x onto $S_1 \cap S_2$ (right-hand side of (1-1)).

2. An example in $\mathbb{R}^2$

To illustrate von Neumann's theorem we consider the $\mathbb{R}^2$ case. Let $a_1, b_1, a_2, b_2 \in \mathbb{R}$ and let

$$S_1 = \{(x, y) \mid a_1 x + b_1 y = 0\} \quad \text{and} \quad S_2 = \{(x, y) \mid a_2 x + b_2 y = 0\}.$$

In order for S_1 and S_2 to be distinct 1-dimensional subspaces, we require that the a_i and b_i are not both zero[1] and that $a_1/b_1 \neq a_2/b_2$. Since the orthogonal projection onto a subspace is a linear transformation, we can represent such transformations by matrices. In the plane, the matrix that projects any point in $\mathbb{R}^2$ onto S_i is given by

$$A_i = \frac{1}{a_i^2 + b_i^2} \begin{pmatrix} b_i^2 & -a_i b_i \\ -a_i b_i & a_i^2 \end{pmatrix},$$

where $i = 1, 2$. Therefore, the matrix $A = A_2 A_1$ gives us the composition of the two projections.

$$A = \frac{a_1 a_2 + b_1 b_2}{(a_1^2 + b_1^2)(a_2^2 + b_2^2)} \begin{pmatrix} b_1 b_2 & -a_1 b_2 \\ -a_2 b_1 & a_1 a_2 \end{pmatrix}$$

To compute iterations of the matrix A, we wish to express A in terms of a diagonal matrix D similar to A. This is possible, of course, if A is nondefective; that is, if the dimension of each of the eigenspaces of A is equal to the multiplicity of the corresponding eigenvalue. It is easily shown that A is nondefective in the $\mathbb{R}^2$ case. The matrix S of eigenvectors of A is then

$$S = \begin{pmatrix} a_1 & b_1 \\ b_1 & -a_2 \end{pmatrix},$$

with D being

$$D = S^{-1} A S.$$

Computing powers of the matrix A is then a matter of raising the eigenvalues of A to that power:

$$A^k = S D^k S^{-1}.$$

Applying von Neumann's theorem to this equation, we obtain

$$\lim_{k \to \infty} (A_2 A_1)^k = \lim_{k \to \infty} A^k = S \left(\lim_{k \to \infty} D^k \right) S^{-1} = A_\infty,$$

where A_∞ is the matrix representation of $P_{S_1 \cap S_2}$. Note that the limit exists if the eigenvalues of A have absolute value less than or equal to unity.

---

[1]If, say, $a_1 = b_1 = 0$ then $S_1 = \mathbb{R}^2$.

3. Solution algorithm

It is possible to extend the solution method in the previous section to $\mathbb{R}^n$. Here we present a brief outline of the solution algorithm, as explained in [Hoffman and Kunze 1971].

(1) Choose bases for S_1 and S_2.

(2) Use the Gram–Schmidt procedure to produce orthonormal bases $\beta^{(1)}$ and $\beta^{(2)}$ for S_1 and S_2 respectively:

$$\beta^{(1)} = \{u_1^{(1)}, \ldots, u_{n-1}^{(1)}\}, \quad \beta^{(2)} = \{u_1^{(2)}, \ldots, u_{n-1}^{(2)}\}. \tag{3-1}$$

(3) Use the standard basis $\beta = \{e_1, \ldots, e_n\}$ for the parent vector space $\mathbb{R}^n$.

(4) Use the following general formula to obtain the matrix representations A_i, with $i = 1, 2$, of the orthogonal projections $P_i : \mathbb{R}^n \to S_i$:

$$A_i = \left[\left(\sum_{j=1}^{n-1} \langle e_1, u_j^{(i)} \rangle u_j^{(i)} \right), \ \ldots, \ \left(\sum_{j=1}^{n-1} \langle e_n, u_j^{(i)} \rangle u_j^{(i)} \right) \right].$$

(5) Compute $A = A_2 A_1$. Find the eigenvalues $\lambda_1, \ldots, \lambda_n$ and corresponding independent eigenvectors $E_1, \ldots, E_n$ of A. These give us the $n \times n$ matrices

$$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}, \quad S = (E_1, \ldots, E_n).$$

(6) Compute S^{-1}.

(7) Iteration now proceeds as follows:

$$\begin{aligned} v_k = A v_{k-1} &= (SDS^{-1}) v_{k-1} = (SDS^{-1})(SDS^{-1}) v_{k-2} \\ &= \cdots = (SD^k S^{-1}) v_0 = A^k v_0 \end{aligned} \tag{3-2}$$

for $k = 1, 2, 3, \ldots$.

(8) Finally, we obtain $v_\infty = [S(\lim_{k \to \infty} D^k) S^{-1}] v_0$.

In step (5), we rely on the assumption that the matrix A is nondefective in order to find a similar diagonal matrix. We address this question in Section 5.

4. Eigenvalues in $\mathbb{R}^3$: geometric argument

If we consider two 2-dimensional subspaces in 3-space, S_1 and S_2, it is easy to illustrate geometrically the eigenvectors of the alternating projections. By examining a picture of two planes containing the origin in $\mathbb{R}^3$, we see three different types of eigenvectors; the first two are trivial, but the third is less so (refer to Figure 1).
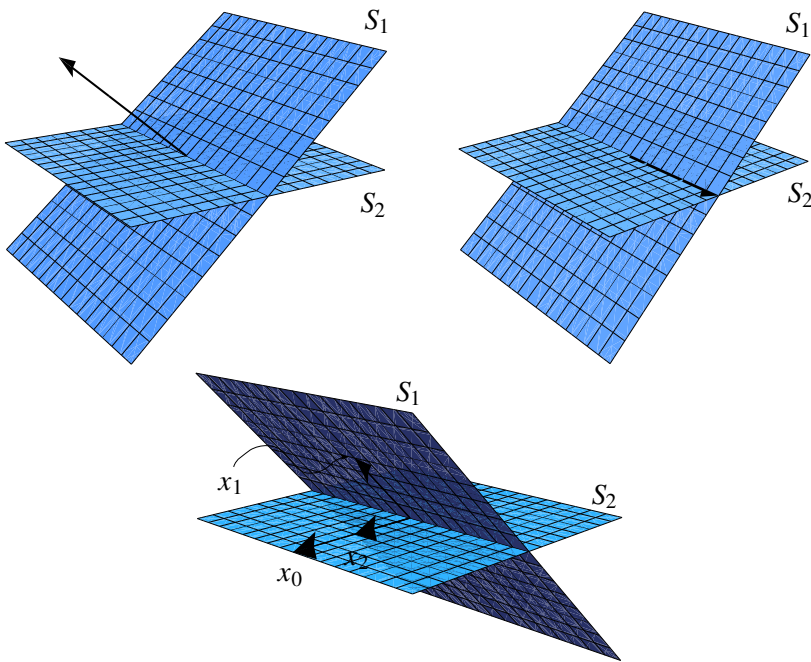
Figure 1. Top left: a vector orthogonal to S_1 gets projected to the origin (eigenvalue 0). Top right: a vector in $S_1 \cap S_2$ remains fixed (eigenvalue 1). Bottom: a vector in $(S_1 \cap S_2)^\perp$ gets projected to a collinear vector (eigenvalue in $[0, 1]$).

(1) A vector orthogonal to S_1 is in the kernel of P_{S_1}; therefore, it is an eigenvector of P_{S_1} with eigenvalue 0.

(2) A vector in $S_1 \cap S_2$ is an eigenvector of both P_{S_2} and P_{S_1} with eigenvalue 1.

(3) A vector in the orthogonal complement $(S_1 \cap S_2)^\perp$ will stay in $(S_1 \cap S_2)^\perp$ as it is projected orthogonally onto S_1 and S_2; i.e., $(S_1 \cap S_2)^\perp$ is invariant under both P_{S_1} and P_{S_2}. Therefore, a vector in $S_2 \cap (S_1 \cap S_2)^\perp$ is an eigenvector of $P_{S_2} P_{S_1}$. We claim that this eigenvector corresponds to an eigenvalue in the interval $[0, 1]$.

It is easy to see from this geometric argument the characterization of eigenvalues in the case of $\mathbb{R}^3$. Next we address the question of whether this geometric intuition somehow generalizes to $\mathbb{R}^n$.

5. Characterization of eigenvalues in $\mathbb{R}^n$.

When we consider $(n-1)$-dimensional subspaces in $\mathbb{R}^n$, it is easy to see that the first two eigenvectors described in Section 4 generalize to higher dimensions. It

is less trivial to show that the third type of eigenvector also generalizes to higher dimensions, and that these three types of vectors fully characterize the spectrum of $P_{S_2}P_{S_1}$.

Let S_1 and S_2 be $(n-1)$-dimensional subspaces of $\mathbb{R}^n$ with $S_1 \neq S_2$.

Lemma 2. $S_1 \cap S_2$ *is a proper subspace of* $\mathbb{R}^n$ *with* $\dim(S_1 \cap S_2) = n - 2$.

Proof. The intersection of two subspaces is always a subspace. Note that for two distinct subspaces, we have

$$n = \dim(S_1) + \dim(S_2) - \dim(S_1 \cap S_2).$$

Therefore,

$$\dim(S_1 \cap S_2) = \dim(S_1) + \dim(S_2) - n$$
$$= n - 1 + n - 1 - n = n - 2. \qquad \square$$

Now, let $S_3 = (S_1 \cap S_2)^\perp$. Note that $n = \dim(S_1 \cap S_2) + \dim(S_3)$, which implies that $\dim(S_3) = 2$.

Lemma 3. $\qquad\qquad \dim(S_3 \cap S_1) = \dim(S_3 \cap S_2) = 1.$

Proof. We write $\dim(S_3 \cap S_1) = \dim(S_3) + \dim(S_1) - n = 2 + n - 1 - n = 1$. Similarly, $\dim(S_3 \cap S_2) = 1$. $\qquad \square$

Lemma 4. *Let* $T_1 : \mathbb{R}^n \to S_1$ *and* $T_2 : \mathbb{R}^n \to S_2$ *be the orthogonal projections onto* S_1 *and* S_2, *respectively. Then* S_3 *is invariant under* T_1 *and* T_2.

Proof. Let $\{w, w^\perp\}$ be a basis for S_3 such that $w \in S_1$ and $w^\perp \in S_1^\perp$. If $v_0 \in S_3$, then $v_0 = c_1 w + c_2 w^\perp$ for some scalars c_1, c_2; therefore,

$$T_1(v_0) = c_1 T_1(w) + c_2 T_1(w^\perp) = c_1 w \in S_3.$$

Similarly, we can construct a basis $\{u, u^\perp\}$ for S_3 such that $u \in S_2$ and $u^\perp \in S_2^\perp$ to conclude that $T_2(v_0) \in S_3$. $\qquad \square$

Now we are ready to prove the following theorem. Let θ be the angle between two hyperplanes defined as the angle between two vectors n_1 and n_2 normal to S_1 and S_2, respectively. Note that $n_1, n_2 \in S_3$.

Theorem 5. *Let* S_1 *and* S_2 *be distinct* $(n-1)$-*dimensional subspaces of* $\mathbb{R}^n$, *and let* $T_1 : \mathbb{R}^n \to S_1$ *and* $T_2 : \mathbb{R}^n \to S_2$ *be the orthogonal projections onto* S_1 *and* S_2, *respectively. Also, let* $0 < \theta < \frac{\pi}{2}$ *be the angle between the two hyperplanes. The spectrum of* $T := T_2 T_1$ *is characterized by the following eigenvalues and multiplicities*:

$$\lambda_1 = 0, \; m_1 = 1, \quad \lambda_2 = 1, \; m_2 = n - 2, \quad \lambda_3 = \cos^2 \theta, \; m_3 = 1.$$

Proof. First, consider u_0 to be a vector orthogonal to S_1. Then $T(u_0) = 0$, and so $m_1 \geq 1$. Now let $\{w_1, \ldots, w_{n-2}\}$ be a basis for $S_1 \cap S_2$. Then $T(w_i) = w_i$ for all $1 \leq i \leq n-2$. Therefore, $\lambda_2 = 1$ is an eigenvalue. Since the basis vectors for $S_1 \cap S_2$ are linearly independent eigenvectors corresponding to λ_2, we have $m_2 \geq n-2$. Furthermore, consider $v_0 \in S_3 \cap S_2$. Then $T(v_0) \in S_3$ by Lemma 4, and $T(v_0) \in S_2$ since the range of T is S_2. Moreover,

$$\dim(S_3 \cap S_2) = 1;$$

therefore, $T(v_0) = \lambda v_0$ for some scalar λ. Furthermore, let $v_1 := T_1(v_0)$ and $v_2 := T_2(v_1) = T(v_0)$. For vectors n_1 and n_2 in the orthogonal complement of S_1 and S_2, respectively, we have that $n_1, n_2, v_0, v_1,$ and v_2 are coplanar, since they are in the 2-dimensional subspace S_3. Thus

$$\angle(v_0, v_1) = \angle(v_1, v_2) = \angle(n_1, n_2) = \theta.$$

Hence, $\cos\theta = \dfrac{\langle v_0, v_1 \rangle}{\|v_0\|\|v_1\|}$ and

$$\|v_2\|\|v_1\|\cos\theta = \frac{\|v_2\|}{\|v_0\|}\langle v_0, v_1 \rangle = \lambda\langle v_0, v_1 \rangle.$$

Note that $\langle v_1, (v_0 - v_1) \rangle = \langle v_2, (v_1 - v_2) \rangle = 0$, so

$$\|v_2\|\|v_1\|\cos\theta = \lambda\langle v_1 + (v_0 - v_1), v_1 \rangle = \lambda\|v_1\|^2 :$$

thus $\dfrac{\|v_2\|}{\|v_1\|}\cos\theta = \lambda$. Moreover,

$$\|v_2\|\|v_1\|\cos\theta = \langle v_2, v_1 \rangle = \langle v_2, v_2 + (v_1 - v_2) \rangle = \|v_2\|^2,$$

so $\cos\theta = \dfrac{\|v_2\|}{\|v_1\|}$. It follows that $\lambda = \cos^2\theta$. $\qquad\square$

6. Conclusion

We have shown that for every finite-dimensional inner product space, the method of alternating orthogonal projections between two hyperplane subspaces S_1 and S_2 yields at most three distinct eigenvalues when we consider the composition of two orthogonal projections. Also, the eigenvectors of such a composition can be quickly identified to be in the subspaces $S_1^\perp$, $S_1 \cap S_2$, and $S_2 \cap (S_1 \cap S_2)^\perp$. We should mention the special, and somewhat trivial, cases where the angle between S_1 and S_2 is $0°$ or $90°$. In the case where $\theta = 90°$, we have that $P_{S_2}P_{S_1} = P_{S_1 \cap S_2}$, and $P_{S_2}P_{S_1} = P_{S_1} = P_{S_2}$ when $\theta = 0°$. In these cases, there are two distinct eigenvalues: 0 and 1. For $\theta = 90°$, the respective multiplicities are 2 and $n-2$; for $\theta = 0°$, they are 1 and $n-1$. It is also noteworthy that the multiplicities obtained in Theorem 5 guarantee that $P_{S_2}P_{S_1}$ is nondefective, a necessary condition for the algorithm presented in Section 3.

Acknowledgements

References

[Deutsch 2001] F. Deutsch, *Best approximation in inner product spaces*, CMS Books in Math. **7**, Springer, New York, 2001. MR 2002c:41001 Zbl 0980.41025

[Hoffman and Kunze 1971] K. Hoffman and R. Kunze, *Linear algebra*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1971. MR 43 #1998 Zbl 0212.36601

rholy2@gmail.com   *Johns Hopkins University, 29 Hillview Avenue,*
                                    Boston, MA 02131, United States

marcolopez@my.unt.edu   *Department of Mathematics, University of North Texas,*
                                    1716 W. Hickory St. Apt 2, Denton, TX 76201, United States

douglas.mupasiri@uni.edu   *Department of Mathematics, University of Northern Iowa,*
                                    220 Wright Hall, Cedar Falls, IA 50614-0506, United States

mln40@msstate.edu   *Jackson State University, 8845 Hwy 12 West,*
                                    Sallis, MS 39160, United States

■■
■■msp
■

The 3-point Steiner problem on a cylinder

Denise M. Halverson and Andrew E. Logan

(Communicated by Frank Morgan)

The 3-point Steiner problem in the Euclidean plane is to find the least length path network connecting three points. In this paper we will demonstrate an algorithm for solving the 3-point Steiner problem on the cylinder.

1. Introduction

Say we have three points on a cylinder. What would be the shortest possible path network connecting our three points? Our goal is to develop an algorithm to find the minimal path network connecting three points on a cylinder. Finding the least length path network connecting a given set of fixed points in a surface is called the Steiner problem. We will first show that the Steiner problem on the cylinder is related to the Steiner problem on the plane. We then will work with a covering map from the plane to the cylinder so that the correspondence between the Steiner problem on the plane and on the cylinder is clarified. We will follow this with a few results culminating in the cutting theorem. The cutting theorem, Theorem 5.3, guarantees that for any configuration of three points on a cylinder there exists a straight line in the cylinder through which we can make a "cut," then flatten the cut surface out in the plane, and finally construct the minimal path network connecting the three points within the flattened surface. The cutting theorem is an important result that leads us to the cutting algorithm. The cutting algorithm determines the minimal path network connecting the three points on the cylinder. The algorithm requires two cuts in order to compare the principal minimal path network candidates obtained when flattening the cut surface of the cylinder out in the plane.

Only within the last 40 years has the Steiner problem really begun to be studied on nonplanar surfaces. Local properties of minimal path networks on smooth surfaces were investigated in [Weng 2001]. Cockayne [1972] and Brazil et al. [1998] provided analytic methods to solve the 3-point Steiner problem in the sphere. Analytic methods for finding the solution to Steiner problems on the hyperbolic plane and surfaces of revolution were given in [Halverson and March 2005] and

[Caffarelli et al. 2012], respectively. Geometric methods for solving the two- and
3-point Steiner problems on the regular tetrahedron were provided in [Brune and
Sipe 2009; Moon et al. 2011]. A cutting algorithm to find the solution to 3-point
Steiner problems on the cone, similar to the one in this paper, is given in [Lee et al.
2011]. Results providing for reductions in solving the 3-point Steiner problem on
the torus are found in [Halverson and Penrod 2007; Ivanov and Tuzhilin 1994;
May and Mitchell 2007]. Furthermore, Ivanov and Tuzhilin [1994] classify all the
closed local minimal networks on closed surfaces of constant nonnegative curvature
(spheres, projective planes, flat tori, and Klein bottles) and present similar results
for the regular tetrahedron. Helmandollar and Penrod [2007] used a generalization
of the method of paired calibrations to solve Steiner problems in the hyperbolic
plane for four fixed points that are the vertices of a square. Hwang et al. [1992]
offer a detailed discussion on various strategies, extensions, and modifications of
the Steiner problem.

The importance of this paper is that it provides an algorithm that does not just
give a reduction to the list of possible solutions or refer to a set of analytic equations
which must be solved, but finds an actual geometric solution to any 3-point Steiner
problem on the cylinder.

2. The Steiner problem on the plane

In this section we will give a brief background of the Steiner problem in the plane.
For a more extensive study on the Steiner problem in the Euclidean plane see
[Hwang et al. 1992; Ivanov and Tuzhilin 1994]. First we will begin with a few
definitions and a basic result concerning the Steiner problem. Then we will give a
brief history of the development of solutions to this problem. Finally, we will finish
with an algorithm for finding a minimal path network connecting three points in
the plane.

Definition 2.1. Let A, B, and C be points in $\mathbb{R}^2$. A *Steiner minimal tree*, denoted
SMT(A, B, C), is the set of minimal length path networks contained in $\mathbb{R}^2$ that
connect A, B, and C.

It is a classical result that, for three points A, B, and C in the plane, SMT(A, B, C)
contains precisely one element (see [Hwang et al. 1992]). It is a common practice
to denote this unique path network itself as SMT(A, B, C). We will also apply
this practice in our paper when considering the 3-point Steiner problem on the
plane. It is also a classical result that, if $\triangle ABC$ has no interior angle with measure
$\geq 120°$, then SMT$(A, B, C) = \overline{AS} \cup \overline{BS} \cup \overline{CS}$ for some point S, called the *Steiner
point* (see [Courant and Robbins 1979]). In this case we say that SMT(A, B, C) is
full. If $\triangle ABC$ has an interior angle with measure $\geq 120°$, say $m\angle ABC \geq 120°$,
then SMT$(A, B, C) = \overline{AB} \cup \overline{BC}$. In this case we say SMT(A, B, C) is *degenerate*.

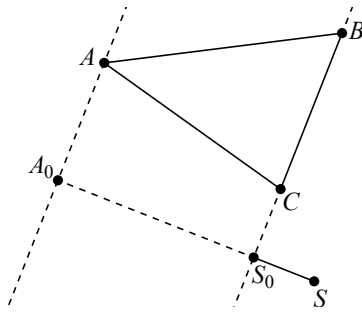**Figure 1.** Demonstrating that τ_0 is shorter than τ in the proof of Propositon 2.3.

Note that in this case $\mathrm{SMT}(A, B, C) = \overline{AB} \cup \overline{BB} \cup \overline{BC}$, so in some sense B takes on a similar role as the Steiner point in the full case.

Definition 2.2. Let A, B, and C be points in $\mathbb{R}^2$. We call the point S a *generalized Steiner point* if $\overline{AS} \cup \overline{BS} \cup \overline{CS} \in \mathrm{SMT}(A, B, C)$.

Another result of the Steiner problem in the plane is that the minimal path network connecting three points in a plane is contained in the convex hull of the triangle whose vertices lie on those three points. Since we use this result in proving future theorems in this paper, we will demonstrate a proof here in this section.

Propositon 2.3. *If A, B, and C are points in the plane, then* $\mathrm{SMT}(A, B, C)$ *is contained in the convex hull of* $\triangle ABC$.

Proof. Let $\tau \in \mathrm{SMT}(A, B, C)$ and let $S \in \mathbb{R}^2$ be the generalized Steiner point of τ.

Suppose τ is not contained in the convex hull of $\triangle ABC$. Then S lies outside of the convex hull of $\triangle ABC$. Hence S is opposite one of the points A, B, or C of the lines $\overleftrightarrow{BC}$, $\overleftrightarrow{AC}$, or $\overleftrightarrow{AB}$, respectively. Suppose without loss of generality S is on the side of the line $\overleftrightarrow{BC}$ opposite point A (see Figure 1). Then there is a line perpendicular to $\overleftrightarrow{BC}$ that passes through S. Let S_0 be the point of intersection of the two lines. Let $\tau_0 = \overline{AS_0} \cup \overline{BS_0} \cup \overline{CS_0}$. Note that $SS_0 > 0$ because S is not on $\overleftrightarrow{BC}$. Since $BS = \sqrt{(BS_0)^2 + (SS_0)^2}$ and $CS = \sqrt{(CS_0)^2 + (SS_0)^2}$, then $BS_0 < BS$ and $CS_0 < CS$. Let l be the line parallel to BC passing through A and let A_0 be the point of intersection of l and $\overleftrightarrow{SS_0}$. Since $A_0S_0 < A_0S$,

$$AS = \sqrt{(AA_0)^2 + (A_0S)^2} > \sqrt{(AA_0)^2 + (A_0S_0)^2} = AS_0.$$

Thus τ_0 is shorter than τ, which yields a contradiction.

Therefore τ is contained in the convex hull of $\triangle ABC$. $\square$

Other interesting results of the Steiner problem on the plane are found in [Cieslik 1998; Hwang et al. 1992; Ivanov and Tuzhilin 1994; Jarník and Kössler 1934; Lee et al. 2011; Roussos 2012].
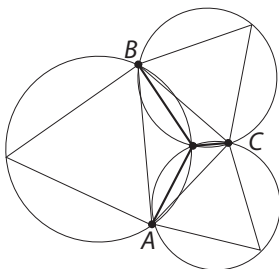
Figure 2. Torricelli's solution.

Brief history. The history of the Steiner problem is briefly described in [Cieslik 1998; Courant and Robbins 1979; Kuhn 1974; Roussos 2012]. We give a summary here.

Fermat posed the following problem in the early 17th century: "Given three points in the plane, find a fourth point such that the sum of its distances to the three given points is minimal." Around 1640 Torricelli presented a geometric solution to Fermat's problem. He showed in the full case that the three circles circumscribing the equilateral triangles constructed on the sides of and outside the triangle intersect at the desired point which is often referred to as the Fermat–Torricelli point [Cieslik 1998] (see Figure 2). SMT(A, B, C) is the configuration of the bold lines in Figure 2. In 1836 Gauss considered the Fermat problem for $n > 3$ points, sometimes referred to as Gauss's problem.

Steiner gave a geometric construction of the Fermat–Torricelli point in the early 19th century and used it in the construction of distance-minimizing trees and graphs [Roussos 2012]. Courant and Robbins [1979] popularized the minimizing of path networks for n points and (mis)labeled it the *Steiner problem*; see [Cieslik 1998] for discussion.

Note that Torricelli's solution only holds when all angles in $\triangle ABC$ are less than or equal to 120°. If we were to perform Torricelli's algorithm of the solution on a triangle with an interior angle greater than 120 degrees we would get a point outside of the convex hull of that triangle which contradicts Propositon 2.3; hence the distinction between full and degenerate minimal path networks.

Solution to the 3-point Steiner problem in the plane. We will now present a useful algorithm [Melzak 1961] for finding SMT(A, B, C) and its length.

First draw the triangle connecting the three points. If one of the angles of $\triangle ABC$ has measure $\geq$ 120°, remove the opposite side. The union of the remaining two sides is SMT(A, B, C) and its length is the sum of the lengths of the two sides. In this case SMT(A, B, C) is degenerate.

Otherwise choose one of the sides of the triangle (for example in Figure 3 we chose side $\overline{BC}$) and draw an equilateral triangle, $\triangle BCE$, where E is on the side of
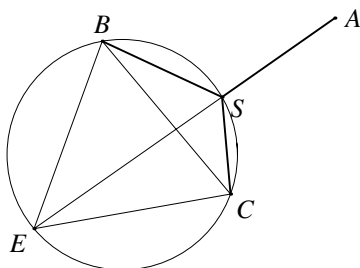
Figure 3. Constructing a full minimal length path network in the plane.

the line $\overline{BC}$ opposite point A. Draw a circle circumscribing $\triangle BCE$ and draw a line from point E to point A. The intersection of the line and the circle will give us the point S, the Steiner point. Then EA will be the length of SMT(A, B, C), and SMT$(A, B, C) = \overline{AS} \cup \overline{BS} \cup \overline{CS}$. In this case SMT$(A, B, C)$ is full.

3. The cylinder

We will now introduce the cylinder and the covering map we will be using in this paper. (Refer to Figure 4.)

Let $\mathscr{C} \subseteq \mathbb{R}^3$ be the cylinder defined by $\mathscr{C} : x^2 + y^2 = 1$. Then $\mathbb{R}^2$ is a covering for $\mathscr{C}$, where $p : \mathbb{R}^2 \to \mathscr{C}$ is the covering map such that $p(u, v) = (\cos u, \sin u, v)$. Let x denote an arbitrary point of $\mathscr{C}$. Let X_i be the point of $p^{-1}(x)$ contained in $[-\pi + 2i\pi, \pi + 2i\pi)$. We denote by (u_X, v_X) the coordinates of an arbitrary point X in $\mathbb{R}^2$.

Definition 3.1. For points $A, B \in \mathbb{R}^2$ where $A = (u_A, v_A)$ and $B = (u_B, v_B)$, the *strip* Σ_{AB} is the set $\Sigma_{AB} = \{(u, v) \in \mathbb{R}^2 \mid u_A \leq u \leq u_B\}$.

In this paper we will order without loss of generality the three fixed points a, b, and c in such a way that $u_{A_0} \leq u_{B_0} \leq u_{C_0}$.

For a 3-point Steiner problem on a cylinder with fixed points a, b, and c, it will be convenient to distinguish the three regions partitioned by the vertical lines for
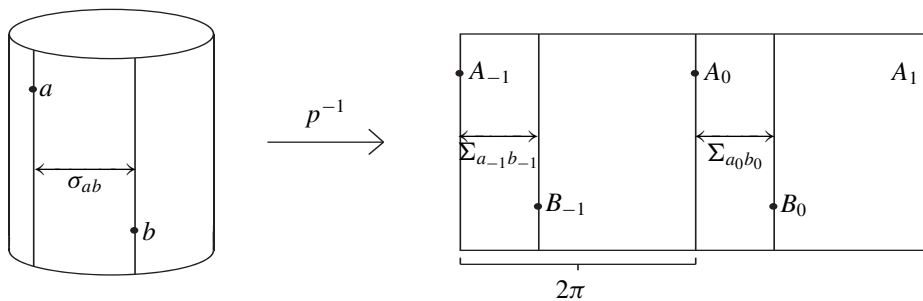

Figure 4. The covering map p.

each of the fixed points. In particular, let $\sigma_{ab} = p(\Sigma_{A_0 B_0})$, $\sigma_{bc} = p(\Sigma_{B_0 C_0})$, and $\sigma_{ca} = p(\Sigma_{C_0 A_0})$.

Definition 3.2. Let $\mathcal{L}$ be a subset of $\mathcal{C}$. A map $f : \mathcal{L} \to \mathbb{R}^2$ is said to be a *lift of the inclusion map* $\mathcal{L} \hookrightarrow \mathcal{C}$ provided, for all $z \in \mathcal{L}$, $z = p \circ f(z)$. We also say the set $f(\mathcal{L})$ is a *lift* of $\mathcal{L}$.

4. Regarding the 3-point Steiner problem on a smooth surface

The Steiner problem on any smooth surface is similar to, but more complicated than, the Steiner problem in the plane [Weng 2001]. In this section we provide definitions and notations for a minimal length path network and a generalized Steiner point on a smooth surface. On a cylinder, and in other smooth surfaces, the minimal length path network need not be unique. Hence we have the following definitions.

Definition 4.1. Let a, b, and c be points in a smooth surface $\mathcal{X}$. Then SMT(a, b, c) is the set of minimal path networks contained in $\mathcal{X}$ that connect a, b, and c.

Definition 4.2. Let a, b, and c be points in a smooth surface $\mathcal{X}$. If $\tau = \overline{as} \cup \overline{bs} \cup \overline{cs} \in$ SMT(a, b, c), then we say that s is a *generalized Steiner point* for τ.

There have been many studies of the Steiner problem on general curved surfaces. We cannot address all results and studies in this paper, but refer the interested reader to [Brazil et al. 1998; Cockayne 1972; Dolan et al. 1991; Ivanov and Tuzhilin 1994; Weng 2001] for more details.

5. The cutting theorem

Our purpose in this paper is to present an algorithm for finding a minimal path network on a cylinder. We first need to prove the cutting theorem that we will use in the cutting algorithm; this result, in short, informs us that any minimal path network on a cylinder will be contained in the union of two of the strips σ_{ab}, σ_{bc}, and σ_{ca}. In preparation for the proof of the cutting theorem we need the following proposition.

Propositon 5.1. *Let T be a minimal path network for three fixed points in the plane such that $p(T) \in$ SMT(a, b, c), S be the generalized Steiner point of T, and $X \in T$ be a fixed point of T such that $p(X) \in \{a, b, c\}$. Then $|u_X - u_S| \leq \pi$.*

Proof. Suppose $|u_X - u_S| > \pi$. By properties of the covering map p there is a point $X_i \in p^{-1}(p(X))$ so that $|u_{X_i} - u_S| \leq \pi$. Then T' obtained by replacing $\overline{XS}$ in T with $\overline{X_i S}$ is a shorter path network where $p(T')$ connects a, b, c. Hence $p(T) \notin$ SMT(a, b, c). This is a contradiction, so $|u_X - u_S| \leq \pi$. $\square$

Corollary 5.2. *Let T be a minimal path network for three fixed points in the plane such that $p(T) \in$ SMT(a, b, c) and S be the generalized Steiner point of T. Then $T \subseteq \Gamma = \{(u, v) \in \mathbb{R}^2 : |u - u_S| \leq \pi\}$.*

Proof. Let $T = \overline{A_l S} \cup \overline{B_m S} \cup \overline{C_n S}$. Since $|u_{A_l} - u_S| \leq \pi$, then $\overline{A_l S} \subset \Gamma$. Likewise $\overline{B_m S}, \overline{C_n S} \subset \Gamma$. Thus $T \subseteq \Gamma$. $\qquad\square$

The following theorem demonstrates that the lift of a minimal path network connecting three points a, b, and c on a cylinder is contained in one of the following:

$$\Sigma_{B_{k-1} A_k} = \Sigma_{B_{k-1} C_{k-1}} \cup \Sigma_{C_{k-1} A_k},$$

$$\Sigma_{C_{k-1} B_k} = \Sigma_{C_{k-1} A_k} \cup \Sigma_{A_k B_k},$$

$$\Sigma_{A_k C_k} = \Sigma_{A_k B_k} \cup \Sigma_{B_k C_k}.$$

A proof of a similar result regarding the flat torus can be found in [Halverson and Penrod 2007].

Theorem 5.3 (cutting theorem). *Let T be a minimal path network for three fixed points in the plane such that $p(T) \in \mathrm{SMT}(a, b, c)$. Then T is contained in one of $\Sigma_{B_{k-1} A_k}$, $\Sigma_{C_{k-1} B_k}$, and $\Sigma_{A_k C_k}$ for some $k \in \mathbb{Z}^+$.*

Proof. Let $T = \overline{A_l S} \cup \overline{B_m S} \cup \overline{C_n S}$. Let $t = \min\{u_{A_l}, u_{B_m}, u_{C_n}\}$. Without loss of generality let $t = u_{A_l}$. By Corollary 5.2, $|u_{A_l} - u_S| \leq \pi$ and $|u_{B_m} - u_S| \leq \pi$. Using $u_{A_l} \leq u_{B_m}$ and the triangle inequality, we have

$$u_{B_m} - u_{A_l} = |u_{B_m} - u_{A_l}| \leq |u_{B_m} - u_S| + |u_{A_l} - u_S| \leq 2\pi.$$

Note that $m \geq l$. Let $m = l + j$ for some $j \in \mathbb{Z}^+$. Then $u_{B_m} = u_{B_l} + 2\pi j \leq 2\pi + u_{A_l}$. Thus

$$0 \leq u_{B_l} - u_{A_l} \leq 2\pi - 2\pi j.$$

This is only possible if j is either 0 or 1. Furthermore, when $j = 1$ equality must occur. In particular, if $j = 1$, then $u_{B_l} = u_{A_l}$ and hence $u_{B_m} = u_{A_{l+1}}$. So either $m = l$ or $m = l + 1$, and in the case $m = l + 1$ necessarily $u_{B_m} = u_{A_{l+1}}$. Similar considerations of C_n yield either $n = l$ or $n = l + 1$, and in the case $n = l + 1$ necessarily $u_{C_n} = u_{A_{l+1}}$.

Case 1. Suppose $m = l$ and $n = l$. Then $T = \mathrm{SMT}(A_l, B_l, C_l)$. By Propositon 2.3, T is in the convex hull of $\triangle A_l B_l C_l$. Thus $T \subset \Sigma_{A_l C_l}$. Letting $k = l$ gives the desired result.

Case 2. Suppose $m = l + 1$ and $n = l$. Then $u_{B_m} = u_{A_{l+1}}$ and hence $u_{B_{m-1}} = u_{A_l}$. Thus T is in the convex hull of $\triangle A_l B_{l+1} C_l$. It follows that $T \subset \Sigma_{A_l B_{l+1}} = \Sigma_{B_l A_{l+1}}$. Letting $k = l + 1$ gives the desired result.

Case 3. Suppose $n = l + 1$. Then $u_{C_n} = u_{A_{l+1}}$. Thus $u_{A_l} = u_{C_l}$. Since $u_{A_l} \leq u_{B_l} \leq u_{C_l}$, then $u_{A_l} = u_{B_l} = u_{C_l}$. Since $v_{X_i} = v_{X_j}$ for any $i, j \in \mathbb{Z}$, the length of T is

$$\sqrt{(u_S - u_{A_{l+1}})^2 + (v_S - v_{A_{l+1}})^2} + \sqrt{(u_S - u_{B_m})^2 + (v_S - v_{B_m})^2} + \sqrt{(u_S - u_{C_n})^2 + (v_S - v_{C_n})^2}$$

$$\geq |v_S - v_{A_{l+1}}| + |v_S - v_{B_m}| + |v_S - v_{C_{l+1}}|$$

$$\geq |v_S - v_{A_l}| + |v_S - v_{B_l}| + |v_S - v_{C_l}|$$

$$\geq \max\{|v_{A_l} - v_{C_l}|, |v_{A_l} - v_{B_l}|, |v_{B_l} - v_{C_l}|\}$$

$$= \max\{A_l C_l, A_l B_l, B_l C_l\}.$$

Let T' be the minimal path connecting A_l, B_l, and C_l. Note that, since A_l, B_l and C_l are collinear, T' is one of $\overline{A_l C_l}$, $\overline{A_l B_l}$, and $\overline{B_l C_l}$. Then the length of T' is $\max\{A_l C_l, A_l B_l, B_l C_l\}$ which is less than or equal to the length of T. Also note that equality can only hold when $u_S = u_{A_l} = u_{B_l} = u_{C_l}$, implying $T = T'$ which is a contradiction. Therefore this case is not possible.

Similar arguments apply when $t = u_{B_m}$ and $t = u_{C_n}$. $\qquad\qquad$ $\square$

6. The cutting algorithm

Justification. Let a, b, and c be points on the cylinder $\mathscr{C}$ and let T be a lift of $\tau \in \mathrm{SMT}(a, b, c)$ contained in $\Sigma_{B_{-1}C_0}$. This is possible from the cutting theorem since we know that there is a lift of τ contained in one of $\Sigma_{B_{-1}A_0}$, $\Sigma_{C_{-1}B_0}$, and $\Sigma_{A_0 C_0}$. Notice that if we cut along the vertical line containing a and lay it out in a plane we get copies of $\Sigma_{B_{-1}A_0}$ and $\Sigma_{A_0 C_0}$, contained in the cut surface. If we cut along the vertical line containing b and lay it out in a plane we get copies of $\Sigma_{B_{-1}A_0}$ and $\Sigma_{C_{-1}B_0}$, contained in the cut surface. If we cut along the vertical line containing c and lay it out in a plane we get copies of $\Sigma_{A_0 C_0}$ and $\Sigma_{C_{-1}B_0}$, contained in the cut surface. With all three cuts together we get copies of each of $\Sigma_{B_{-1}A_0}$, $\Sigma_{C_{-1}B_0}$, and $\Sigma_{A_0 C_0}$ twice. One way to determine the $\mathrm{SMT}(a, b, c)$ is comparing the minimal path networks in each strip. However the following algorithm demonstrates how to do this more efficiently with just two cuts.

The cutting algorithm. **Step 1.** Cut along the vertical line containing a of our cylinder. Then there are two possible minimal path networks, one in $\Sigma_{A_0 C_0}$ and one in $\Sigma_{B_{-1}A_0}$. Let T_1 be $\mathrm{SMT}(A_0, B_0, C_0)$, and T_2 be $\mathrm{SMT}(B_{-1}, C_{-1}, A_0)$. Since T_1 and T_2 are both in the plane, perform the algorithm presented in Section 2 to compare the two minimal path networks and find which one is shorter.

Step 2. If T_1 is at least as short as T_2, then cut vertically up the cylinder at the point c and unwrap it as before, laying it out on the plane contained in $\Sigma_{C_{-1}C_0}$. Then there are two possible minimal path networks, one in $\Sigma_{C_{-1}B_0}$ and the other in $\Sigma_{A_0 C_0}$. Let T_3 be $\mathrm{SMT}(C_{-1}, A_0, B_0)$. Note that T_1 is contained in $\Sigma_{A_0 C_0}$. Since T_3 is in the plane, use the algorithm for finding minimal path networks in the plane

presented in Section 2 and compare T_3 to T_1. Let i be any index where T_i is at least as short as T_j for all $j \neq i$. Then $p(T_i) \in \mathrm{SMT}(a, b, c)$.

Otherwise, cut vertically up the cylinder at the point b and unwrap it, laying it out on the plane contained in $\Sigma_{B_{-1}B_0}$. Then there are two possible minimal path networks, one in $\Sigma_{B_{-1}A_0}$ and the other in $\Sigma_{C_{-1}B_0}$. Let T_3 be $\mathrm{SMT}(C_{-1}, A_0, B_0)$ as in the first case. Note that T_2 is contained in $\Sigma_{B_{-1}A_0}$. Since T_3 is in the plane use the algorithm for finding minimal path networks in the plane presented in Section 2 and compare T_3 to T_2. Let i be any index where T_i is at least as short as T_j where $j \neq i$. Then $p(T_i) \in \mathrm{SMT}(a, b, c)$.

That's all there is to it.

7. Conclusion

Further problems that can be investigated include:

(1) The n-point Steiner problem on the cylinder. Jarník and Kössler [1934] have developed an algorithm for solving any n-point Steiner problem in the plane. How could the results in this paper be generalized to solve any n-point problem on the cylinder?

(2) The 3-point Steiner problem on the flat torus in 4-space. The cylinder is a covering space for the flat torus in 4-space. How can the results produced in this paper be applied to solve the 3-point Steiner problem on the flat torus in 4-space?

(3) The n-point Steiner problem on the flat torus in 4-space. Could results of (1) and (2) be combined to solve any n-point Steiner problem on the flat torus in 4-space?

We hope that the results found in this paper can serve as a basis in many future findings.

References

[Brazil et al. 1998] M. Brazil, J. H. Rubinstein, D. A. Thomas, J. F. Weng, and N. C. Wormald, "Shortest networks on spheres", pp. 453–461 in *Network design: Connectivity and facilities location* (Princeton, NJ, 1997), edited by P. M. Pardalos and D. Du, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. **40**, Amer. Math. Soc., Providence, RI, 1998. MR 98k:05046 Zbl 0915.05043

[Brune and Sipe 2009] T. Brune and L. Sipe, "Shortest path between two points on the regular tetrahedron", preprint, 2009, http://tinyurl.com/BruneSipe2009.

[Caffarelli et al. 2012] E. A. Caffarelli, D. M. Halverson, and R. J. Jensen, "The Steiner problem on surfaces of revolution", *Graphs and Combinatorics* (2012). To appear.

[Cieslik 1998] D. Cieslik, *Steiner minimal trees*, Nonconvex Optimization and its Applications **23**, Kluwer Academic Publishers, Dordrecht, 1998. MR 99i:05062 Zbl 0997.05500

[Cockayne 1972] E. J. Cockayne, "On Fermat's problem on the surface of a sphere", *Math. Mag.* **45** (1972), 216–219. MR 47 #4145 Zbl 0257.52013

[Courant and Robbins 1979] R. Courant and H. Robbins, *What is mathematics?: An elementary approach to ideas and methods*, Oxford University Press, New York, 1979. MR 80i:00001 Zbl 0442.00001

[Dolan et al. 1991] J. Dolan, R. Weiss, and J. M. Smith, "Minimal length tree networks on the unit sphere", *Ann. Oper. Res.* **33**:1-4 (1991), 503–535. MR 92k:68098 Zbl 0741.90081

[Halverson and March 2005] D. Halverson and D. March, "Steiner tree constructions in hyperbolic space", preprint, 2005, http://tinyurl.com/HalversonMarch2005.

[Halverson and Penrod 2007] D. Halverson and K. Penrod, "Three-point Steiner problem on the flat torus", preprint, 2007, http://tinyurl.com/HalversonPenrod2007.

[Helmandollar and Penrod 2007] H. Helmandollar and K. Penrod, "Length minimizing paths in the hyperbolic plane: Proof via paired subcalibrations", *Illinois J. Math.* **51**:3 (2007), 723–729. MR 2009m:51029 Zbl 1146.51014

[Hwang et al. 1992] F. K. Hwang, D. S. Richards, and P. Winter, *The Steiner tree problem*, Annals of Discrete Mathematics **53**, North-Holland Publishing Co., Amsterdam, 1992. MR 94a:05051 Zbl 0774.05001

[Ivanov and Tuzhilin 1994] A. O. Ivanov and A. A. Tuzhilin, *Minimal networks: The Steiner problem and its generalizations*, CRC Press, Boca Raton, FL, 1994. MR 95h:05050 Zbl 0842.90116

[Jarník and Kössler 1934] V. Jarník and M. Kössler, "O minimálních grafech, obsahujících *n* daných bodů", *Časopis pro pěstování matematiky a fysiky* **63**:8 (1934), 223–235. Zbl 0009.13106

[Kuhn 1974] H. W. Kuhn, ""Steiner's" problem revisited", pp. 52–70 in *Studies in optimization*, edited by G. B. Dantzig and B. C. Eaves, Studies in Math. **10**, Math. Assoc. Amer., Washington, D.C., 1974. MR 57 #18835 Zbl 0347.90054

[Lee et al. 2011] A. Lee, J. Lytle, D. Halverson, and D. Sampson, "The Steiner problem on narrow and wide cones", preprint, 2011, http://tinyurl.com/aq4qdfd.

[May and Mitchell 2007] K. L. May and M. A. Mitchell, "The three point Steiner problem on the flat torus: The minimal lune case", preprint, 2007, http://tinyurl.com/MayMitchell2007.

[Melzak 1961] Z. A. Melzak, "On the problem of Steiner", *Canad. Math. Bull.* **4** (1961), 143–148. MR 23 #A2767 Zbl 0101.13201

[Moon et al. 2011] K. Moon, G. Shero, and D. Halverson, "The Steiner problem on the regular tetrahedron", *Involve* **4**:4 (2011), 365–404. MR 2905235 Zbl 1245.51005

[Roussos 2012] I. M. Roussos, "On the Steiner minimizing point and the corresponding algebraic system", *College Math. J.* **43**:4 (2012), 305–308. MR 2974509

[Weng 2001] J. F. Weng, "Steiner trees on curved surfaces", *Graphs Combin.* **17**:2 (2001), 353–363. MR 2003c:05058 Zbl 0982.05036

halverson@math.byu.edu   *Department of Mathematics, Brigham Young University, Provo, UT 84602, United States*

aelogan13@gmail.com   *Department of Mathematics, Brigham Young University, Provo, UT 84602, 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.