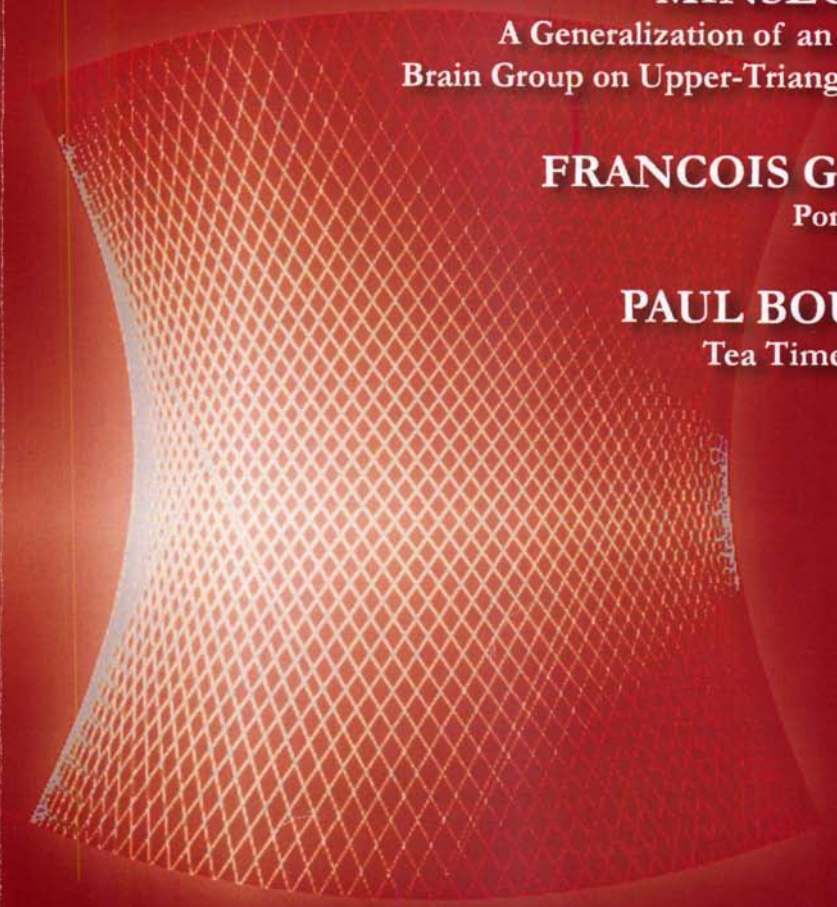
The Harvard College Mathematics Review

Volume 4

Spring 2012

In this issue:



HCMR

A Student Publication of Harvard College

Website. Further information about The HCMR can be found online at the journal's website,

http://www.thehcmr.org/ (1)

Instructions for Authors. All submissions should include the name(s) of the author(s), institutional affiliations (if any), and both postal and e-mail addresses at which the corresponding author may be reached. General questions should be addressed to Editor-In-Chief Rediet Abebe at hcmr@hcs.harvard.edu.

Articles. The Harvard College Mathematics Review invites the submission of quality expository articles from undergraduate students. Articles may highlight any topic in undergraduate mathematics or in related fields, including computer science, physics, applied mathematics, statistics, and mathematical economics.

Authors may submit articles electronically, in .pdf, .ps, or .dvi format, to hcmr@hcs.harvard.edu, or in hard copy to

The Harvard College Mathematics Review
Student Organization Center at Hilles
Box # 360
59 Shepard Street
Cambridge, MA 02138.

Submissions should include an abstract and reference list. Figures, if used, must be of publication quality. If a paper is accepted, high-resolution scans of hand drawn figures and/or scalable digital images (in a format such as .eps) will be required.

Problems. The HCMR welcomes submissions of original problems in all mathematical fields, as well as solutions to previously proposed problems.

Proposers should send problem submissions to Problems Editor Yale Fan at hcmr-problems@hcs.harvard.edu or to the address above. A complete solution or a detailed sketch of the solution should be included, if known.

Solutions should be sent to hcmr-solutions@hcs.harvard.edu or to the address above. Solutions should include the problem reference number. All correct solutions will be acknowledged in future issues, and the most outstanding solutions received will be published.

Advertising. Print, online, and classified advertisements are available; detailed information regarding rates can be found on The HCMR's website, (1). Advertising inquiries should be directed to hcmr-advertise@hcs.harvard.edu, addressed to Business Manager Hamsa Sridhar.

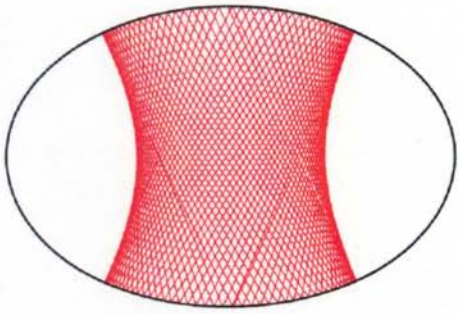
Subscriptions. One-year (two issue) subscriptions are available, at rates of $5.00 for students, $7.50 for other individuals, and $15.00 for institutions. Subscribers should mail checks for the appropriate amount to The HCMR's postal address; confirmation e-mails or queries should be directed to hcmr-subscribe@hcs.harvard.edu.

Sponsorship. Sponsoring The HCMR supports the undergraduate mathematics community and provides valuable high-level education to undergraduates in the field. Sponsors will be listed in the print edition of The HCMR and on a special page on the The HCMR's website, (1). Sponsorship is available at the following levels:

Sponsor	$0 - $99
Fellow	$100 - $249
Friend	$250 - $499
Contributor	$500 - $1,999
Donor	$2,000 - $4,999
Patron	$5,000 - $9,999
Benefactor	$10,000 +

Contributors · The Harvard University Mathematics Department

Cover Image. The image on the cover depicts an integral billiard (ellipse). This issue's article "Tea Time in Princeton" by Dr. Paul Bourgade (p. 41) discusses eigenvalues of random matrices and their connection to number theory.



The Harvard College Mathematics Review is produced and edited by a student organization of Harvard College.

-2
Contents

-1
Staff

Editors-In-Chief
Rediet Abebe '13 and Greg Yang '14

Articles Editor
Geoffrey Lee '14
Problems Editor
Yale Fan '14

Features Editor
Lucia Mocz '13
Business Manager
Hamsa Sridhar '12

Issue Production Director
Eric Larson '13

Graphic Artist
Keoni Correa '13

Editors Emeritus
Zachary Abel '10, Ernest Fontes '10, Scott Kominers '09/AM'10/PhD'11

Webmasters
Rediet Abebe '13 and Akhil Mathew '14

Board of Reviewers
Rediet Abebe '13
Levent Alpoge '14
John Casale '12
Irving Dai '14
Yale Fan '14
Marco Gentili '15
Eric Larson '13
Geoffrey Lee '14
Akhil Mathew '14
Lucia Mocz '13
Seth Neel '15
Toan Phan '14
Caitlin Stanton '14
Allen Yuan '15

Faculty Adviser
Professor Peter Kronheimer, Harvard University

0 From the Editors

Rediet Abebe
Harvard University '13
Cambridge, MA 02138
rtesfaye@college.harvard.edu

Greg Yang
Harvard University '14
Cambridge, MA 02138
gyang@college.harvard.edu

Many of us have grown up with math books and math competitions. We remember the adrenaline rush as the clock ticked down on the tests and the sleepless nights spent on attacking problems and deciphering theories. We still maintain the friendships that are made through the common interest of mathematics. That part of our childhoods have long been ingrained into our characters, and we inherit these symptoms of love for mathematics even in college.

Now we, the HCMR staff, have the honor to celebrate this love by furnishing the texts that would hopefully become the material of somebody's sleepless nights. For this process, we have had the fortune of many people's support. The two of us would like to thank the **HCMR staff** and **executive members** for the endless enthusiasm and dedication to the making of this journal. It has been a privilege to work with such committed members over the past year. The HCMR could not have grown to the level of visibility across departments around the nation and the world without your contribution.

As always, we owe our deepest gratitude to members of the **Harvard College Math Department** for the advice, encouragement and guidance throughout the publication process. We would like to thank the HCMR's **advisors** and **sponsors** for the profound contributions which continue to make this journal a success; and the **Student Organization Center at Hilles** for providing the HCMR a space in which to grow.

Finally, we would like to thank our **readers** for all the input you have given us in person and via email. Please feel free to continue to contact us with any comments, questions and concerns you may have.

This issue marks the fourth volume since the inception of the Harvard College Math Review and the second since it became an annual publication. Over the five years that the HCMR has stood as a student-run organization, it has had celebration of mathematics both in its pure and applied form at the heart of its mission. We hope you find joy in such a celebration as we did in creating this journal.

Rediet Abebe '13 and Greg Yang '14
Editors-In-Chief, The HCMR

STUDENT ARTICLE

1

A Generalization of an Action of the Braid Group on Upper-Triangular Matrices

Minseon Shin†
Massachusetts Institute of Technology '13
Cambridge, MA 02138
mshin@mit.edu

Abstract

We investigate various questions related to a conjecture posed by A. I. Bondal in his 2004 paper "A symplectic groupoid of triangular bilinear forms." Bondal describes an action of the braid group of n strings $\mathbf{B}_n$ on the $n \times n$ upper-triangular matrices, which has a natural (linear) algebraic formulation obtained by embedding $\mathbf{B}_n$ into a group of matrices whose entries are polynomials in $\frac{n(n-1)}{2}$ indeterminates (with an unusual composition law). One can try to extend the action to other matrices in this group to get a larger group acting on the set of upper-triangular matrices. Bondal conjectures that the action cannot enlarge past $\mathrm{Im}(\mathbf{B}_n \ltimes (\mathbb{Z}/2\mathbb{Z})^n)$. In this paper, we prove Bondal's conjecture for the case $n = 2$, study the elements that act trivially, and present an unfinished inductive strategy for proving the conjecture for $n \geq 3$.

1.1 Introduction

Let k be an algebraically closed field of characteristic zero, let V be a vector space of dimension n over k, and let $\langle\ ,\ \rangle$ be a nondegenerate bilinear form on V. Let $\mathbb{E}$ be the set of ordered bases of V.

Definition 1. An ordered basis $E = (e_1, \ldots, e_n)$ is called *semiorthogonal* if $\langle e_i, e_j \rangle = 0$ for all $i > j$ and $\langle e_i, e_i \rangle = 1$ for all i.

Let $\mathbb{E}_s \subset \mathbb{E}$ be the set of ordered semiorthogonal bases of V and consider the group $\mathrm{Perm}(\mathbb{E}_s)$ of permutations of $\mathbb{E}_s$. Bondal proved in [1] that $\mathrm{Perm}(\mathbb{E}_s)$ contains a subgroup isomorphic to the braid group, which is defined as follows.

Definition 2. The *braid group* $\mathbf{B}_n$ has generators $\{\sigma_1, \ldots, \sigma_{n-1}\}$ with relations

$$\begin{aligned} \sigma_i\sigma_j &= \sigma_j\sigma_i && \text{if } |i-j| > 1 \\ \sigma_i\sigma_{i+1}\sigma_i &= \sigma_{i+1}\sigma_i\sigma_{i+1} && \text{for } 1 \leq i \leq n-2\,. \end{aligned}$$

Definition 3. For $i = 1, \ldots, n-1$, define the transformation $\varphi_i : \mathbb{E} \to \mathbb{E}$ which maps the ordered basis $E = (e_1, \ldots, e_n)$ to $\varphi_i(E) = (e_1', \ldots, e_n')$ where

$$\begin{aligned} e_i' &= e_{i+1} - \langle e_i, e_{i+1} \rangle e_i \\ e_{i+1}' &= e_i \\ e_j' &= e_j \text{ for } j \notin \{i, i+1\}\,. \end{aligned}$$

†Minseon Shin, Massachusetts Institute of Technology '13, is a mathematics major. His current mathematical interests include algebra and combinatorics. Outside of academics, he enjoys listening to music and playing soccer.

It is easily checked that if E is semiorthogonal, then $\varphi_i(E)$ is semiorthogonal. Therefore φ_i permutes the set of semiorthogonal bases: $\varphi_i \in \mathrm{Perm}(\mathbb{E}_s)$. The transformation φ_i, when restricted to $\mathbb{E}_s$, has an inverse $\varphi_i^{-1} : \mathbb{E}_s \to \mathbb{E}_s$ which maps $E = (e_1, \ldots, e_n)$ to $\varphi_i^{-1}(E) = (e_1', \ldots, e_n')$ where

$$\begin{aligned} e_i' &= e_{i+1} \\ e_{i+1}' &= e_i - \langle e_i, e_{i+1} \rangle e_{i+1} \\ e_j' &= e_j \text{ for } j \notin \{i, i+1\} . \end{aligned}$$

Proposition 4 (Bondal [1] 2.1). *The correspondence $\sigma_i \longmapsto \varphi_i$ can be extended to an action of the braid group by automorphisms on the set of semiorthogonal bases.*

In other words, the correspondence $\sigma_i \longmapsto \varphi_i$ can be extended to a homomorphism $\mathbf{B}_n \to \mathrm{Perm}(\mathbb{E}_s)$. As Bondal notes, this action of $\mathbf{B}_n$ on $\mathbb{E}_s$ is not faithful, because the center $Z(\mathbf{B}_n)$ acts trivially on $\mathbb{E}_s$ (we prove this fact in Section 1.4). However, it is possible to embed this group in a particular group of matrices with a special law of composition.

The matrix of the bilinear form with respect to a semiorthogonal basis is upper-triangular with ones on the main diagonal; the space of such upper-triangular matrices with ones on the main diagonal constitutes an affine space of dimension $\frac{n(n-1)}{2}$, which we denote by $\mathcal{A}$.

Let R be a ring, let $\mathbf{X} = \{X_{ij} \ : \ 1 \leq i < j \leq n\}$ be a set of $\frac{n(n-1)}{2}$ indeterminates, let $R[\mathbf{X}]$ be the polynomial ring over the indeterminates $\mathbf{X}$ and coefficients in R, and let $M_n(\mathbf{X}, R)$ be the set of $n \times n$ matrices whose entries are in $R[\mathbf{X}]$. Let X be the $n \times n$ upper-triangular matrix with ones on the main diagonal and whose (i,j)th entry is X_{ij} for all $i < j$. For any $B \in M_n(\mathbf{X}, k)$ and $A \in \mathcal{A}$, we denote by B_A or $B(A)$ the matrix obtained by evaluating B at A; specifically, substituting $A_{ij} \to X_{ij}$.

Let us associate, to each generator σ_i of the braid group $\mathbf{B}_n$, a matrix $\sigma_i(X) \in M_n(\mathbf{X}, k)$ which coincides with the identity matrix I_n at all entries except for the 2×2 matrix

$$\begin{bmatrix} 0 & 1 \\ 1 & X_{i,i+1} \end{bmatrix}$$

which is situated so that $X_{i,i+1}$ is the $(i+1, i+1)$th entry of $\sigma_i(X)$. If $n = 3$, for example, we have

$$\sigma_1(X) = \begin{bmatrix} & 1 & \\ 1 & X_{12} & \\ & & 1 \end{bmatrix} \quad \text{and} \quad \sigma_2(X) = \begin{bmatrix} 1 & & \\ & & 1 \\ & 1 & X_{23} \end{bmatrix} .$$

The motivation for this definition is the following property of semiorthogonal bases. For every $E \in \mathbb{E}_s$, let A_E be the matrix of the form $\langle \ , \ \rangle$ with respect to E. Then A_E is upper-triangular and $\sigma_i(A_E)$ is the change-of-basis matrix from $\varphi_i(E)$ to E; in other words, the matrix which satisfies $E = \varphi_i(E)\sigma_i(A_E)$. Since $\sigma_i(X)$ satisfies $\varphi_i(E) = E\sigma_i(A_E)^{-1}$, the matrix of the form $\langle \ , \ \rangle$ with respect to the basis $\varphi_i(E)$ is $A' = \sigma_i(A_E)^{-T} A_E \sigma_i(A_E)^{-1}$, which is contained in $\mathcal{A}$ since φ_i preserves semiorthogonality on $\mathbb{E}$. It is easily verified that, in fact, $\sigma_i(A)^{-T} A \sigma_i(A)^{-1} \in \mathcal{A}$ for all $A \in \mathcal{A}$.

We now consider the set of all elements of $M_n(\mathbf{X}, k)$ that have the above property.

Definition 5. Let $\mathcal{B}_0(n)$ be the set of elements $B \in M_n(\mathbf{X}, k)$ such that B has a two-sided inverse B^{-1} with entries in the fractional field of $k[\mathbf{X}]$ and satisfying

$$B_A^{-T} A B_A^{-1} \in \mathcal{A} \text{ for all } A \in \mathcal{A} . \tag{1.1}$$

The condition that some $B \in M_n(\mathbf{X}, k)$ is contained in $\mathcal{B}_0(n)$ is equivalent to requiring that, if A is the matrix of the form of a semiorthogonal basis E, then the new basis $E' = EB_A^{-1}$ should also be semiorthogonal. Notice that requiring $B_A^{-T} A B_A^{-1} \in \mathcal{A}$ for all $A \in \mathcal{A}$ implies that $B_X^{-T} X B_X^{-1}$ itself is upper-triangular, since the lower-left entries of $B_X^{-T} X B_X^{-1}$ are rational functions in the indeterminates $\mathbf{X}$ which vanish on all of $\mathcal{A}$.

Definition 6. $\mathcal{B}_0(n)$ is a monoid with the composition law $*$ defined as follows: For $B, C \in \mathcal{B}_0(n)$, let

$$B * C = B_{C_X^{-T} X C_X^{-1}} C_X \ . \tag{1.2}$$

This law of composition $*$ is well-defined, associative, and has an identity which is simply the identity matrix I (these properties may be verified easily).

Definition 7. Let us denote by $\mathcal{B}(n) \subset \mathcal{B}_0(n)$ the set of elements in $\mathcal{B}_0(n)$ which have two-sided inverses with respect to $*$.

Then $\sigma_i(A) \in \mathcal{B}(n)$ for all i, where the inverse of $\sigma_i(A)$ with respect to $*$ is $\sigma_i^{-1}(A)$. (In general, $\sigma_i^{-1}(A)$ is *not* equal to $(\sigma_i(A))^{-1}$, the inverse of $\sigma_i(A)$ with respect to matrix multiplication.)

There are other elements of $M_n(\mathbf{X}, k)$ that, trivially, lie in $\mathcal{B}(n)$. For every $s = (s_1, \ldots, s_n) \in (\mathbb{Z}/2\mathbb{Z})^n$, define the diagonal matrix N_s whose (i,i)th entry is $(-1)^{s_i}$, which represents the basis transformation $(e_1, \ldots, e_n) \longmapsto \left((-1)^{s_1} e_1, \ldots, (-1)^{s_n} e_n\right)$. Then N_s lies in $\mathcal{B}(n)$, and its inverse with respect to $*$ is itself.

Definition 8. There is a natural action $\cdot : \mathcal{B}(n) \times \mathcal{A} \to \mathcal{A}$ of $\mathcal{B}(n)$ on $\mathcal{A}$, defined as follows:

$$B \cdot A = B_A^{-T} A B_A^{-1} \ . \tag{1.3}$$

Proposition 9 (Bondal [1] pg. 669). *Consider the semidirect product $\mathbf{B}_n \ltimes (\mathbb{Z}/2\mathbb{Z})^n$; the function*

$$\varphi : \mathbf{B}_n \ltimes (\mathbb{Z}/2\mathbb{Z})^n \to \mathcal{B}(n) \tag{1.4}$$

mapping $\sigma_i \longmapsto \sigma_i(A)$ for $\sigma_i \in \mathbf{B}_n$ and $s \longmapsto N_s$ for $s \in (\mathbb{Z}/2\mathbb{Z})^n$ is a monomorphism.

We are primarily concerned with the following conjecture posed by Bondal:

Conjecture 10 (Bondal [1] 2.2). *The monomorphism φ defined in Proposition 9 is also surjective.*

Bondal stated that he knew the proof for the cases $n \leq 3$. While we do not prove the conjecture in its entirety, we describe and prove a number of small related results.

Structure of paper. In Section 1.2, we state and prove some basic constraints on an element B of $\mathcal{B}(n)$. In Section 1.3, we prove Conjecture 10 for the case $n = 2$. In Section 1.4, we study elements of $\mathcal{B}(n)$ acting trivially on $\mathcal{A}$. In (1.4.1), we explicitly compute the matrix associated to the generator of the center of $\mathbf{B}_n$. Though we cannot prove that this generates, up to signs, the group of elements in $\mathcal{B}(n)$ acting trivially, we prove some lemmas toward this (1.4.2). In Section 1.5, we present an inductive strategy for proving Conjecture 10 for higher values of n; we do not fulfill this agenda, but set up some of the steps.

Sections 1.3, 1.4, and 1.5 each require the reader to have read only Sections 1.1 and 1.2. Personally, I feel the most interesting results and conjectures are Lemma 18, Conjecture 20, the discussion in Section 1.4.1, and Proposition 44.

Acknowledgements. This work was undertaken as part of the Summer Program for Undergraduate Research run by the MIT math department over summer 2011. I would like to thank the MIT math department for supporting me, Professor David Jerison for organizing SPUR, Professor Paul Seidel for suggesting the problem, and my mentor Ailsa Keating for her mathematical insights and advice, as well as her extensive help on editing and formatting this paper.

1.2 Basic Constraints

1.2.1 Coordinate-free constraints

Lemma 11 and its Corollary 12 show that if $B \in \mathcal{B}_0(n)$ then the entries of B^{-1} are actually elements of $k[\mathbf{X}]$, instead of rational functions in $k(\mathbf{X})$, as originally defined in Definition 5. Proposition 13 is stated without proof since it not used elsewhere in this paper; however, it may provide the reader with intuition on the structure of $\mathcal{B}_0(n)$.

Lemma 11. *Suppose that* $B \in \mathcal{B}_0(n)$. *Then* $\det B = \pm 1$.

Proof. If $B \in \mathcal{B}_0(n)$, then $A' = B_A^{-T} A B_A^{-1} \in \mathcal{A}$ for all $A \in \mathcal{A}$. Since $\det A = 1$ for all $A \in \mathcal{A}$, we have $(\det B_A)^{-2} = \det(B_A^{-T} A B_A^{-1}) = \det A' = 1$ which implies $\det B_A = \pm 1$. Since B_A depends continuously on A, so does $\det B_A$, so $\det B = +1$ (or -1) identically. □

Corollary 12. *Suppose that* $B \in \mathcal{B}_0(n)$. *Then* B^{-1} *has polynomial entries.* □

Proposition 13. *(a) If* $B \in \mathcal{B}_0(n)$, *then* B_I *is orthogonal. Moreover,* $B \cdot I = I$.

(b) If $B \in \varphi(\mathbf{B}_n)$, *then* B_I *is a permutation matrix.*

(c) If $B \in \mathcal{B}_0(n) \cap M_n(\mathbf{X}, \mathbb{Z})$, *then* B_I *is a permutation matrix (up to signs).*

(d) If $B \in \operatorname{Im}\varphi$, *then* $B \in M_n(\mathbf{X}, \mathbb{Z})$.

1.2.2 A Symmetric Polynomial Equation

Suppose that $B \in \mathcal{B}(n)$; we investigate a polynomial equation satisfied by the coordinates of each column of B^{-1}. Definition 14 is motivated by Lemma 15.

Definition 14. For any $\mathbf{f} := (f_1, \ldots, f_n) \in (k[\mathbf{X}])^n$ and any nonempty $S \subset [n]$, define

$$P_S(\mathbf{f}) := \sum_{i \in S} f_i^2 + \sum_{i,j \in S, i<j} X_{ij} f_i f_j \, .$$

Lemma 15. *Suppose that* $B \in \mathcal{B}(n)$ *and* b_{ij} *is the* (i,j)*th entry of* B^{-1}. *Fix* $1 \le \ell \le n$ *and let* $f_i = b_{i\ell} \in k[\mathbf{X}]$ *for all* i. *Then* $\mathbf{f} = (f_1, \ldots, f_n)$ *satisfies the relation* $P_{[n]}(\mathbf{f}) = 1$.

Proof. Let $[f_1 \ \ldots \ f_n]^T$ be the ℓth column of B_X^{-1}. Since $B_X^{-T} X B_X^{-1} \in \mathcal{A}$, we have

$$\begin{aligned} P_{[n]}(\mathbf{f}) &= \sum_{i=1}^{n} f_i^2 + \sum_{1 \le i < j \le n} X_{ij} f_i f_j \\ &= [f_1 \ \ldots \ f_n] \begin{bmatrix} 1 & X_{12} & \cdots & X_{1n} \\ & 1 & \cdots & X_{2n} \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix} \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix} \\ &= (B_X^{-T} X B_X^{-1})_{\ell\ell} \\ &= 1. \end{aligned}$$

□

By Lemma 15, any column of B^{-1} is a solution to $P_{[n]}(\mathbf{f}) = 1$. Therefore, we are interested in the set of solutions $\mathbf{f}$ to $P_{[n]}(\mathbf{f}) = 1$.

The equation $P_{[n]}(\mathbf{f}) = 1$ is quadratic when considered as a polynomial in f_1. Since quadratic equations have at most two distinct solutions, the set of solutions f to $P_{[n]}(\mathbf{f}) = 1$ is "discrete" in the sense that if $f_2, \ldots, f_n$ are fixed, then there are at most two polynomials f_1 which satisfy $P_{[n]}(\mathbf{f}) = 1$.

In order to describe the set of such solutions $\mathbf{f} = (f_1, \ldots, f_n)$ to $P_{[n]}(\mathbf{f}) = 1$, we consider the following transformation.

Definition 16. For any $\ell \in [n]$, define the transformation $T_\ell : (k[\mathbf{X}])^n \to (k[\mathbf{X}])^n$ which maps $(f_1, \ldots, f_n)$ to $(f_1', \ldots, f_n')$ where

$$f_i' = \begin{cases} -f_i & \text{if } i \neq \ell \\ f_\ell + \sum_{j \neq \ell} X_{j\ell} f_j & \text{if } i = \ell \end{cases}.$$

Lemma 17. *For any* ℓ, *the composition* $T_\ell \circ T_\ell$ *is the identity.* □

Lemma 18. *For any ℓ, the transformation T_ℓ preserves the quantity $P_{[n]}(\mathbf{f})$. In other words, $P_{[n]}(T_\ell(\mathbf{f})) = P_{[n]}(\mathbf{f})$.*

Proof. We prove the lemma for $\ell = n$; the proof is analogous for $\ell = 1, \ldots, n-1$. Let $\mathbf{f} := (f_1, \ldots, f_n)$ and $\mathbf{f}' := T_\ell(\mathbf{f}) = (f_1', \ldots, f_n')$. Then

$$\begin{aligned} P_{[n]}(\mathbf{f}') &= P_{[n-1]}(\mathbf{f}') + f_n' \left(f_n' + \sum_{i=1}^{n-1} X_{in} f_i' \right) \\ &= P_{[n-1]}(\mathbf{f}) + \left(f_n + \sum_{i=1}^{n-1} X_{in} f_i \right) f_n \\ &= P_{[n]}(\mathbf{f}) \end{aligned}$$

□

Let $\mathbf{Y} = \{Y_i : 1 \le i \le m\}$ be another set of indeterminates. In Lemma 19 and Conjecture 20, we consider the f_i to be elements in $k[\mathbf{X}, \mathbf{Y}]$ instead of in $k[\mathbf{X}]$ in the hope that it will help in an inductive solution, for example in Proposition 44.

Lemma 19. *Suppose that $\mathbf{f} = (f_1, \ldots, f_n) \in (k[\mathbf{X}, \mathbf{Y}])^n$ satisfies $P_{[n]}(\mathbf{f}) = 0$ or 1. Let d_i be the (total) degree of f_i and let $d = \max_i\{d_i\}$ and suppose that $d \ge 1$. If there is exactly one ℓ such that $d_\ell = d$, then T_ℓ reduces the degree of f_ℓ.*

Proof. Let $d_\ell = d$. The condition that there is exactly one ℓ such that $d_\ell = d$ implies that $d_i < d_\ell$ if $i \neq \ell$. If $d_i < d_\ell - 1$ for all $i \neq \ell$, then f_ℓ^2 has the unique highest degree in $P_{[n]}(\mathbf{f})$, contradiction. Thus there exists some i such that $d_i = d_\ell - 1$. Let S be the set of all indices i such that $d_i = d_\ell - 1$. Let $g_i \in k[\mathbf{X}, \mathbf{Y}]$ be the "leading term" of f_i, the sum of monomials of maximal total degree in f_i. Then $g_\ell^2 + \sum_{i \in S} X_{i\ell} g_i g_\ell = 0$. But $g_\ell + \sum_{i \in S} X_{i\ell} g_i$ is the leading term of the ℓth term of $T_\ell(f)$. □

Conjecture 20. *All the solutions to $P_{[n]}(\mathbf{f}) = 1$ for $f_1, \ldots, f_n \in k[\mathbf{X}, \mathbf{Y}]$ can be reduced by the transformations T_ℓ to one of the "elementary solutions", in which all but one f_i are zero and the nonzero f_i is either ± 1. More precisely, for any solution f to $P_{[n]}(\mathbf{f}) = 1$, there exists a finite sequence $a_1, \ldots, a_N \in [n]$ such that $T_{a_1}(\cdots(T_{a_N}(f))\cdots)$ is an elementary solution.*

1.3 The case $n = 2$

We give a full proof of Conjecture 10 for the case $n = 2$. We use coordinates, so the lemmas given here depend heavily on the the fact that $n = 2$. We have not been able to generalize most of them to higher dimensions.

We will use the notation:

$$X = \begin{bmatrix} 1 & x \\ & 1 \end{bmatrix} \quad \text{and} \quad B_X = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

where $b_{ij} \in k[x]$. Let $\beta := \det B$. Then

$$X^{-1} = \begin{bmatrix} 1 & -x \\ & 1 \end{bmatrix}; \quad \sigma_1(X) = \begin{bmatrix} & 1 \\ 1 & x \end{bmatrix}; \quad (\sigma_1(X))^{-1} = \begin{bmatrix} -x & 1 \\ 1 & \end{bmatrix}; \quad \sigma_1^{-1}(X) = \begin{bmatrix} x & 1 \\ 1 & \end{bmatrix}.$$

We start with two technical lemmas that will get used in the proof the theorem.

Lemma 21. *Let $B \in M_2(\mathbf{X}, k)$; then $B \in \mathcal{B}_0(2)$ if and only if the entries b_{ij} satisfy the following conditions:*

$$b_{21}^2 + b_{22}^2 - x b_{21} b_{22} = 1 \tag{1.5}$$

$$b_{11}^2 + b_{12}^2 - x b_{11} b_{12} = 1 \tag{1.6}$$

$$x b_{12} b_{21} - b_{11} b_{21} - b_{12} b_{22} = 0 \tag{1.7}$$

Proof. This is immediate from

$$B_X^{-T} X B_X^{-1} = \begin{bmatrix} b_{21}^2 + b_{22}^2 - x b_{21} b_{22} & x b_{11} b_{22} - b_{11} b_{21} - b_{12} b_{22} \\ x b_{12} b_{21} - b_{11} b_{21} - b_{12} b_{22} & b_{11}^2 + b_{12}^2 - x b_{11} b_{12} \end{bmatrix}. \qquad \square$$

Lemma 22. *If $B \in \mathcal{B}_0(2)$, then*

(a) $B_X^{-T} X B_X^{-1} = \begin{bmatrix} 1 & \beta x \\ & 1 \end{bmatrix}$.

(b) $\begin{bmatrix} b_{22} & b_{21} \\ b_{12} & b_{11} \end{bmatrix} \in \mathcal{B}_0(2)$.

(c) $b_{21} = -\beta b_{12}$.

(d) $b_{22} = \beta(-x b_{12} + b_{11})$.

(e) Suppose that $b_{ij} \neq 0$ for all i, j and let $d_{ij} = \deg b_{ij}$. Then there exists a positive integer m such that one of the following holds:

$$\{d_{11} = m - 1\ ,\ d_{12} = d_{21} = m\ ,\ d_{22} = m + 1\}$$
$$\{d_{11} = m + 1\ ,\ d_{12} = d_{21} = m\ ,\ d_{22} = m - 1\}.$$

Proof. Each of these properties follows from manipulating the conditions of Lemma 21; we omit the proofs. $\square$

We are now ready for the proof of Conjecture 10 for $n = 2$. We proceed in two steps.

Lemma 23. $\mathcal{B}_0(2) = \mathcal{B}(2)$. *Given $B \in \mathcal{B}_0(2)$, an explicit inverse with respect to $*$ is C given by:*

$$C = \begin{cases} B_X^{-1} & \text{if } \det B = 1 \\ B_{X^{-1}}^{-1} & \text{if } \det B = -1\ . \end{cases}$$

Proof. Suppose $B \in \mathcal{B}_0(2)$ and $\det B = -1$; the case $\det B = 1$ is similar, and we omit it. Then $B_X^{-T} X B_X^{-1} = X^{-1}$ if and only if $B_X^T X^{-1} B_X = X$. Let $C := B_{X^{-1}}^{-1}$; then $C \in M_2(\mathbf{X}, k)$ and $C \cdot X = B_{X^{-1}}^T X B_{X^{-1}} = X^{-1} \in \mathcal{A}$ so $C \in \mathcal{B}_0(2)$. It suffices now to show that $B * C = C * B = I$. Since $C \cdot X = X^{-1}$, we have $B * C = B_{C \cdot X} C_X = B_{X^{-1}} C_X = I$. Also, $B \cdot X = X^{-1}$ so $C * B = C_{B \cdot X} B_X = C_{X^{-1}} B_X = I$. $\square$

Theorem 24. *Every $B \in \mathcal{B}(2)$ is contained in the image of $\mathbf{B}_2 \ltimes (\mathbb{Z}/2\mathbb{Z})^2$.*

Proof. We proceed case by case. Suppose $b_{12} = 0$ then, by Lemma 22(c), $b_{21} = 0$. By Lemma 21 (1.5,1.6), $b_{11}^2 = b_{22}^2 = 1$. This gives us the four matrices $\left[\begin{smallmatrix} \pm 1 & 0 \\ 0 & \pm 1 \end{smallmatrix}\right]$, which is exactly the image of $(\mathbb{Z}/2\mathbb{Z})^2$.

Suppose that $b_{12}, b_{21} \neq 0$. If $b_{11} = 0$, then (1.6) implies that $b_{12} = \pm 1$. So $b_{21} = \pm 1$ by Lemma 22(c). Now (1.7) implies that $b_{22} = b_{21} x$; this is equal to the image of $\sigma_1 s$ for some $s \in (\mathbb{Z}/2\mathbb{Z})^2$. The case $b_{22} = 0$ is similar.

Now let us suppose that all b_{ij} are nonzero. Let m be the integer provided by Lemma 22(e). By Lemma 22(b), we can WLOG assume $\{d_{11} = m - 1, d_{12} = d_{21} = m, d_{22} = m + 1\}$. Let $C = \sigma_s^{-1}(X)$; then $B \in \mathcal{B}(2)$ if and only if $C * B \in \mathcal{B}(2)$. Using Lemma 22(d), we have

$$C * B = C_{B \cdot X} B_X = \begin{bmatrix} \beta x & 1 \\ 1 & \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} \beta x b_{11} + b_{21} & \beta b_{11} \\ b_{11} & b_{12} \end{bmatrix}.$$

If $\beta x b_{11} + b_{21} = 0$, then we are done. Otherwise, $\beta x b_{11} + b_{21}$ has degree $m - 2$ by Lemma 22(e) and we conclude by induction on m. $\square$

1.4 Trivial actions and the center of the braid group

1.4.1 A proof that $\varphi((\sigma_1\sigma_2\cdots\sigma_{n-1})^n) = X^{-T}X$

The center of the braid group B_n is cyclic with generator $(\sigma_1\sigma_2\cdots\sigma_{n-1})^n$ (see e.g. [3]). We prove that the generator $(\sigma_1\sigma_2\cdots\sigma_{n-1})^n$ maps to $X^{-T}X$ under φ, stated without proof in [1].

Lemma 25. *The braid relations imply that*

$$\begin{aligned} \sigma_i(\sigma_1\sigma_2\cdots\sigma_{n-1}) &= (\sigma_1\sigma_2\cdots\sigma_{n-1})\sigma_{i-1} \qquad \text{for} \quad 2 \le i \le n-1 \\ \sigma_i(\sigma_{n-1}\sigma_{n-2}\cdots\sigma_1) &= (\sigma_{n-1}\sigma_{n-2}\cdots\sigma_1)\sigma_{i+1} \qquad \text{for} \quad 1 \le i \le n-2\,. \end{aligned}$$ □

Lemma 26. *Let* $s_n = \sigma_{n-1}\sigma_{n-2}\cdots\sigma_2\sigma_1$. *Then* $s_n^n = (\sigma_1\sigma_2\cdots\sigma_{n-1})^n$.

Proof. By induction. For $n = 2$, there is nothing to prove. For general n we have

$$\begin{aligned} (\sigma_1\sigma_2\cdots\sigma_{n-1})^n &= (\sigma_1\sigma_2\cdots\sigma_{n-2})^{n-1}(\sigma_{n-1}\sigma_{n-2}\cdots\sigma_1)(\sigma_1\sigma_2\cdots\sigma_{n-1}) \\ &= (\sigma_{n-2}\sigma_{n-3}\cdots\sigma_1)^{n-1}(\sigma_{n-1}\sigma_{n-2}\cdots\sigma_1)(\sigma_1\sigma_2\cdots\sigma_{n-1}) \\ &= (\sigma_{n-1}\sigma_{n-2}\cdots\sigma_1)(\sigma_{n-1}\sigma_{n-2}\cdots\sigma_2)^{n-1}(\sigma_1\sigma_2\cdots\sigma_{n-1}) \\ &= (\sigma_{n-1}\sigma_{n-2}\cdots\sigma_1)^n \end{aligned}$$

The second equality follows from the induction hypothesis and the others follow from Lemma 1. □

We prove $\varphi(s_n^n)_X = X^{-T}X$ by expressing both sides as the product of the same elementary matrices, not at first in the same order, then achieve equality by proving that some pairs of these matrices commute.

Notation 27. Let $X_1 = [X_{12} \;\cdots\; X_{1n}]$ and let $\tilde{X}$ be the $(n-1)\times(n-1)$ minor obtained from X by deleting the first row and column, so that $X = \begin{bmatrix} 1 & X_1 \\ & \tilde{X} \end{bmatrix}$. Let I_{n-1} be the $(n-1)\times(n-1)$ identity matrix.

Lemma 28. *Let* L_{ij} *(resp.* L_{ij}^{-1}*) be the elementary matrix which differs from the identity by a* X_{ij} *(resp.* $-X_{ij}$*) in its* (i,j)*th entry. Then*

$$\begin{aligned} X &= (L_{n-1,n})\cdots(L_{23}\cdots L_{2n})(L_{12}\cdots L_{1n}) \\ X^{-T} &= \left(L_{n-1,n}^{-T}\right)\cdots\left(L_{23}^{-T}\cdots L_{2n}^{-T}\right)\left(L_{12}^{-T}\cdots L_{1n}^{-T}\right)\,. \end{aligned}$$

Example 29. For $n = 3$,

$$X = \begin{bmatrix} 1 & & \\ & 1 & X_{23} \\ & & 1 \end{bmatrix}\begin{bmatrix} 1 & X_{12} & \\ & 1 & \\ & & 1 \end{bmatrix}\begin{bmatrix} 1 & & X_{13} \\ & 1 & \\ & & 1 \end{bmatrix}.$$

Proof of Lemma 28. By induction on n. For $n = 2$ there is nothing to prove. Let X_1 and $\tilde{X}$ be defined as in Notation 27. By the induction hypothesis, we have $\begin{bmatrix} 1 & \\ & \tilde{X} \end{bmatrix} = (L_{n-1,n})\cdots(L_{23}\cdots L_{2n})$. Also, we have $\begin{bmatrix} 1 & X_1 \\ & I_{n-1} \end{bmatrix} = L_{12}\cdots L_{1n}$. Combining these two results,

$$X = \begin{bmatrix} 1 & X_1 \\ & \tilde{X} \end{bmatrix} = \begin{bmatrix} 1 & \\ & \tilde{X} \end{bmatrix}\begin{bmatrix} 1 & X_1 \\ & I_{n-1} \end{bmatrix} = (L_{n-1,n})\cdots(L_{23}\cdots L_{2n})(L_{12}\cdots L_{1n})\,.$$ □

Lemma 30. *If* $i_1 < j_1 \le i_2 < j_2$, *then* $L_{i_1j_1}^{-T}$ *commutes with* $L_{i_2j_2}$.

Proof. Let E_{ij} be the $n \times n$ matrix which differs from the zero matrix by a 1 in its (i,j)th entry. Then $E_{i_1j_1}E_{j_2i_2} = E_{j_2i_2}E_{i_1j_1} = 0$ since $j_1 \neq j_2$ and $i_1 \neq i_2$. Thus we have

$$L_{i_1j_1}^{-T} L_{i_2j_2} = (I - X_{i_1j_1}E_{j_1i_1})(I + X_{i_2j_2}E_{i_2j_2}) = I - X_{i_1j_1}E_{j_1i_1} + X_{i_2j_2}E_{i_2j_2}$$
$$L_{i_2j_2} L_{i_1j_1}^{-T} = (I + X_{i_2j_2}E_{i_2j_2})(I - X_{i_1j_1}E_{j_1i_1}) = I + X_{i_2j_2}E_{i_2j_2} - X_{i_1j_1}E_{j_1i_1} ,$$

hence $L_{i_1j_1}^{-T} L_{i_2j_2} = L_{i_2j_2} L_{i_1j_1}^{-T}$. □

Lemma 31. *If $B = \varphi(s_n)$, then*

$$B = \begin{bmatrix} & I_{n-1} \\ 1 & X_1 \end{bmatrix} \quad \textit{and} \quad B \cdot X = \begin{bmatrix} \tilde{X} & -X_1^T \\ & 1 \end{bmatrix} .$$

Example 32. For the case $n = 3$ we have

$$\varphi(\sigma_2\sigma_1)_X = \begin{bmatrix} & 1 & \\ & & 1 \\ 1 & X_{12} & X_{13} \end{bmatrix} \qquad \varphi(\sigma_2\sigma_1) \cdot X = \begin{bmatrix} 1 & X_{23} & -X_{12} \\ & 1 & -X_{13} \\ & & 1 \end{bmatrix} .$$

Proof of Lemma 31. We prove by induction that $\varphi(\sigma_m \cdots \sigma_1)$ transforms the basis

$$E = (e_1, \ldots, e_n) \to E' = (e_1', \ldots, e_n'),$$

where

$$\begin{aligned} e_i' &= e_{i+1} - \langle e_1, e_{i+1}\rangle e_1 \text{ for } i = 1, \ldots, m, \\ e_{m+1}' &= e_1, \text{ and} \\ e_i' &= e_i \text{ for } i > m+1 . \end{aligned}$$

The base case $m = 1$ is immediate. For general m, the basis $E' = (e_1', \ldots, e_n')$ under the transformation $\varphi(\sigma_{m+1}) = \sigma_{m+1}(X)$ is mapped to $E'' = (e_1'', \ldots, e_n'')$ where

$$\begin{aligned} e_{m+1}'' &= e_{m+2}' - \langle e_{m+1}', e_{m+2}'\rangle e_{m+1}' = e_{m+2} - \langle e_1, e_{m+2}\rangle e_1 \\ e_{m+2}'' &= e_{m+1}' = e_1 \\ e_i'' &= e_i' \text{ if } i \neq m, m+1 . \end{aligned}$$

Now the matrix B is the unique element of $M_n(\mathbf{X}, k)$ which satisfies $E'' = EB_{A_E}^{-1}$ for all $E \in \mathbb{E}_s$, thus

$$B^{-1} = \begin{bmatrix} -X_1 & 1 \\ I_{n-1} & \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} & I_{n-1} \\ 1 & X_1 \end{bmatrix}$$

and we have

$$B \cdot X = B_X^{-T} X B_X^{-1} = \begin{bmatrix} -X_1^T & I_{n-1} \\ 1 & \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ & \tilde{X} \end{bmatrix} \begin{bmatrix} -X_1 & 1 \\ I_{n-1} & \end{bmatrix} = \begin{bmatrix} \tilde{X} & -X_1^T \\ & 1 \end{bmatrix} . \qquad \square$$

Lemma 33. *Let $m > 0$. Repeated applications of Lemma 31 show that $\varphi(s_n^m) \cdot X$ is the $(m+1, m+n) \times (m+1, m+n)$ submatrix of $\begin{bmatrix} X & -X^T \\ 0 & X \end{bmatrix}$ and the last row of $\varphi(s_n)_{\varphi(s_n^m)\cdot X}$ is*

$$[1 \; X_{m,m+1} \; \cdots \; X_{mn} \; -X_{1m} \; \cdots \; -X_{m-1,m}] . \qquad \square$$

Example 34. The case $n = 3$. We have already computed $\varphi(\sigma_2\sigma_1)_X$ and $\varphi(\sigma_2\sigma_1)\cdot X$.

$$\varphi(\sigma_2\sigma_1)_{\varphi(\sigma_2\sigma_1)\cdot X} = \begin{bmatrix} 1 & & \\ & & 1 \\ 1 & X_{23} & -X_{12} \end{bmatrix} \qquad \varphi((\sigma_2\sigma_1)^2)\cdot X = \begin{bmatrix} 1 & -X_{13} & -X_{23} \\ & 1 & X_{12} \\ & & 1 \end{bmatrix}$$

$$\varphi(\sigma_2\sigma_1)_{\varphi((\sigma_2\sigma_1)^2)\cdot X} = \begin{bmatrix} 1 & & \\ & & 1 \\ 1 & -X_{13} & -X_{23} \end{bmatrix} \qquad \varphi((\sigma_2\sigma_1)^3)\cdot X = \begin{bmatrix} 1 & X_{12} & X_{13} \\ & 1 & X_{23} \\ & & 1 \end{bmatrix} = X.$$

Lemma 35. *Let P be the permutation matrix which has ones in the entries $(i, i+1)$ for all i (modulo n). If $m > 0$, then*

$$P^{n-m}\varphi(s_n)_{\varphi(s_n^m)\cdot X}P^{m-1} = \left(L_{1m}^{-T}L_{2m}^{-T}\cdots L_{m-1,m}^{-T}\right)\left(L_{m,m+1}L_{m,m+2}\cdots L_{mn}\right).$$

Proof. By Lemma 33, $P^{n-m}\varphi(s_n)_{\varphi(s_n^m)\cdot X}P^{m-1}$ differs from the identity matrix only in the mth row, which is

$$[-X_{1m}\ \cdots\ -X_{m-1,m}\ 1\ X_{m,m+1}\ \cdots\ X_{mn}]\,. \qquad \square$$

Proposition 36. $\varphi(s_n^n)_X = X^{-T}X$.

Example 37. For the case $n = 3$ we have

$$\begin{aligned}
\varphi((\sigma_2\sigma_1)^3)_X &= \left(\varphi(\sigma_2\sigma_1)_{\varphi((\sigma_2\sigma_1)^2)\cdot X}\right)\left(\varphi(\sigma_2\sigma_1)_{\varphi(\sigma_2\sigma_1)\cdot X}\right)\left(\varphi(\sigma_2\sigma_1)_X\right) \\
&= \begin{bmatrix} 1 & & \\ & & 1 \\ 1 & -X_{13} & -X_{23} \end{bmatrix}\begin{bmatrix} 1 & & \\ & & 1 \\ 1 & X_{23} & -X_{12} \end{bmatrix}\begin{bmatrix} 1 & & \\ & & 1 \\ 1 & X_{12} & X_{13} \end{bmatrix} \\
&= \begin{bmatrix} 1 & & \\ & 1 & \\ -X_{13} & -X_{23} & 1 \end{bmatrix}\begin{bmatrix} 1 & & \\ -X_{12} & 1 & X_{23} \\ & & 1 \end{bmatrix}\begin{bmatrix} 1 & X_{12} & X_{13} \\ & 1 & \\ & & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & & \\ -X_{12} & & 1 \\ -X_{13}+X_{12}X_{23} & -X_{23} & 1 \end{bmatrix}\begin{bmatrix} 1 & X_{12} & X_{13} \\ & 1 & X_{23} \\ & & 1 \end{bmatrix} \\
&= X^{-T}X\,.
\end{aligned}$$

Proof of Proposition 36. In the expressions below, the product of matrices is taken with $i = 0$ on the right to $i = n-1$ on the left.

$$\begin{aligned}
\varphi(s_n^n)_X &\overset{1}{=} \prod_{i=0}^{n-1}\varphi(s_n)_{\varphi(s_n^i)\cdot X} \\
&\overset{2}{=} \prod_{i=0}^{n-1}\left(P^{n-1-i}\varphi(s_n)_{\varphi(s_n^i)\cdot X}P^i\right) \\
&\overset{3}{=} \prod_{i=1}^{n}\left(L_{1i}^{-T}L_{2i}^{-T}\cdots L_{i-1,i}^{-T}\right)\left(L_{i,i+1}L_{i,i+2}\cdots L_{in}\right) \\
&\overset{4}{=} \left(\left(L_{n-1,n}^{-T}\right)\cdots\left(L_{23}^{-T}\cdots L_{2n}^{-T}\right)\left(L_{12}^{-T}\cdots L_{1n}^{-T}\right)\right) \\
&\quad\cdot\left(\left(L_{n-1,n}\right)\cdots\left(L_{23}\cdots L_{2n}\right)\left(L_{12}\cdots L_{1n}\right)\right) \\
&\overset{5}{=} X^{-T}X\,.
\end{aligned}$$

Equality 1 follows from the definition of the law of composition $*$; equality 2 follows from the fact that $P^{i+1}P^{n-1-i} = I$ for all $i = 1, \ldots, n-1$; equality 3 follows directly from Lemma 35; equality 4 follows from rearranging the L_{ij} and L_{ij}^{-T} according the commutativity rules in Lemma 30; finally, equality 5 follows directly from Lemma 28. □

1.4.2 Regarding the converse

We now explore the converse: if $B \in \mathcal{B}(n)$ and that B acts trivially on $\mathcal{A}$, then is it true that B is contained in the image of the center of B_n, i.e., $B_X = (X^{-T}X)^k$ for some $k \in \mathbb{Z}$?

To show $B_A = (A^{-T}A)^n$ for all $A \in \mathcal{A}$, it is enough for it to be true in some nonempty open subset of the affine space $\mathcal{A}$.

Lemma 38. *If $B \in \mathcal{B}(n)$ and $B_A^{-T}AB_A^{-1} = A$ for all $A \in \mathcal{A}$, then $B_X(X^{-T}X) = (X^{-T}X)B_X$.*

Proof. Observe that $B_A^{-T}AB_A^{-1} = A$ for all $A \in \mathcal{A}$ if and only if $B_X^{-T}XB_X^{-1} = X$. If this holds, $X^{-T}X = (B_X^{-1}X^{-T}B_X^{-T})(B_X^TXB_X) = B_X^{-1}(X^{-T}X)B_X$. □

We prove that the set of $A \in \mathcal{A}$ such that $A^{-T}A$ has distinct eigenvalues is open and nonempty. Let $c_A(t) = \det(tI - A^{-T}A)$ be the characteristic polynomial of $A^{-T}A$, which is "reciprocal": $c_A(t) = t^n c_A(1/t)$. The condition that $A^{-T}A$ has distinct eigenvalues is equivalent to the condition that the discriminant Δ_A of $c_A(t)$ is nonzero. Since Δ_A is a polynomial in the coefficients of $c_A(t)$, which are themselves polynomials in the variables $\{A_{ij}\}$, we can also express Δ_A as a polynomial in $\{A_{ij}\}$. The subset of $\mathcal{A}$ which satisfies $\Delta_A = 0$ is closed. We now exhibit one A such that $\Delta_A \neq 0$, to show that its complement is nonempty.

Lemma 39. *Let $A \in \mathcal{A}$ be such that $A_{ij} = 1$ for all $i < j$. Then the characteristic polynomial $c_A(t)$ of $A^{-T}A$ is $c_A(t) = \frac{1-(-t)^{n+1}}{1+t}$; in particular, $A^{-T}A$ has distinct eigenvalues.*

Proof. We have

$$(A^{-1})_{ij} = \begin{cases} 1 & \text{if } i = j \\ -1 & \text{if } i = j-1 \\ 0 & \text{otherwise} \end{cases} \implies (A^{-T}A)_{ij} = \begin{cases} 1 & \text{if } i = 1 \\ -1 & \text{if } i = j+1 \\ 0 & \text{otherwise} \end{cases}.$$

Note that $A^T(A^{-T}A)A^{-T} = AA^{-T}$, which is the negative of the companion matrix of the polynomial $\frac{t^{n+1}-1}{t-1}$ (see [2]). Thus the characteristic polynomial of $A^{-T}A$ is $c_A(t) = \frac{(-t)^{n+1}-1}{-t-1}$, whose roots are distinct in any field extension of $\mathbb{Q}$ because $c_A(t)$ and $c_A'(t)$ are relatively prime. □

Let us some fix $A \in \mathcal{A}$ such that $A^{-T}A$ has distinct eigenvalues; then $A^{-T}A$ is diagonalizable. Let P be an invertible matrix such that $D := P^{-1}(A^{-T}A)P$ is diagonal; since B_A and $A^{-T}A$ commute, $P^{-1}B_AP$ and D commute. The only matrices that commute with diagonal matrices with distinct diagonal entries are diagonal matrices, by [4], so $P^{-1}B_AP$ is diagonal. Since the eigenvalues of $A^{-T}A$ are distinct, the Vandermonde matrix of $A^{-T}A$ is invertible, so the powers of D from $i = 0$ to $n-1$ span the subspace of diagonal matrices. Thus there exists a polynomial

$$p(t) = a_{n-1}t^{n-1} + \ldots + a_1t + a_0$$

such that $B_A = p(A^{-T}A)$. Then the condition $B_A^TAB_A = A$ implies

$$p(A^{-T}A)^TAp(A^{-T}A) \iff p(A^{-1}A^T)p(A^{-T}A) = I\ ,$$

therefore $p(D^{-1})p(D) = I$. If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $A^{-T}A$, then $p(\lambda_i^{-1})p(\lambda_i) = 1$ for all i and $p(t^{-1})p(t) \equiv 1 \pmod{c_A}$. In addition, the residue of p is a unit in the quotient ring $k[t, \frac{1}{t}]/\langle c_A\rangle$. As $p(\lambda_i) \neq 0$ for all i, $p(t)$ and $c_A(t)$ are relatively prime.

Lemma 40. *Let $R = k\left[t, \frac{1}{t}\right]$. The units of R are ct^n for $c \in k$ and $n \in \mathbb{Z}$.*

Proof. Let $S = k[t]$ and $f_1, f_2 \in R$ such that $f_1(t)f_2(t) = 1$ for all t. There exists a positive integer N such that $t^N f_1(t)$ and $t^N f_2(t)$ are polynomials (in other words, contained in S). Then $\left(t^N f_1(t)\right)\left(t^N f_2(t)\right) = t^{2N}$ but S is a unique factorization domain, so $t^N f_1(t)$ and $t^N f_2(t)$ are monomials. □

One way to continue this approach would be to try to prove there exists a polynomial $p \in (k[\mathbf{X}])[t]$ such that $B_X = p(X^{-T}X)$.

1.5 Towards an Inductive Proof

Notice that if $C \in \operatorname{Im}\varphi$ and $B \in \mathcal{B}(n)$, then $B \in \operatorname{Im}\varphi$ if and only if $B * C \in \operatorname{Im}\varphi$. Our goal was to prove Conjecture 10 by induction on n; we know it holds for $n = 2$. Our hope was, starting with an element $B \in \mathcal{B}(n)$, to have a method of repeatedly composing it with some generators $\sigma_i(X)$ to get it to the form $B = \begin{bmatrix} \widehat{B} & \\ & 1 \end{bmatrix}$. If we know that $\widehat{B} \in \mathcal{B}(n-1)$, we could proceed inductively.

Lemma 41 shows that it suffices to reduce B so that either the last row or last column has only one nonzero entry, which is necessarily ± 1. We have no conjectures as to the method of reduction. Our induction step is Proposition 44, which assumes Conjecture 42, which we have not been able to prove in full generality. However, we do know how to prove it under the assumption that any $B \in \mathcal{B}(n)$ permutes the elements of $\mathcal{A}$ with entries in $\mathbb{Q}$.

From a geometric background and the motivation of the project (which we do not discuss in this paper), Braid group elements take matrices in $\mathcal{A}$ with integer entries to matrices in $\mathcal{A}$ with integer entries; therefore it seems reasonable to hope that they take matrices in $\mathcal{A}$ with rational entries to matrices in $\mathcal{A}$ with rational entries, although we are not able to prove it.

Lemma 41. *Suppose $B \in \mathcal{B}_0(n)$ is of the form $B = \begin{bmatrix} \widehat{B} & B_1 \\ B_2 & 1 \end{bmatrix}$ where $\widehat{B}$ is an $(n-1) \times (n-1)$ matrix, B_1 is an $(n-1) \times 1$ column vector, and B_2 is a $1 \times (n-1)$ row vector. Then $B_1 = 0$ if and only if $B_2 = 0$.*

Proof. In both directions, for the given form of B, we compute $B \cdot A$ in terms of $\widehat{B}$ and B_i, then use the fact that $B \cdot A \in \mathcal{A}$ to show that $B_{i+1} = 0$.

We first prove that if $B_2 = 0$ then $B_1 = 0$. Suppose $B_2 = 0$; then

$$B^{-1} = \begin{bmatrix} \widehat{B}^{-1} & -\widehat{B}^{-1}B_1 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad B^{-T} = \begin{bmatrix} \widehat{B}^{-T} & 0 \\ -B_1^T\widehat{B}^{-T} & 1 \end{bmatrix},$$

and it follows that

$$B_A^{-T}AB_A^{-1} = \begin{bmatrix} \widehat{B}^{-T}\widehat{A}\widehat{B}^{-1} & (\widehat{B}^{-T}\widehat{A}\widehat{B}^{-1})B_1 + \widehat{B}^{-T}A_1 \\ -B_1^T(\widehat{B}^{-T}\widehat{A}\widehat{B}^{-1}) & -B_1^T(\widehat{B}^{-T}\widehat{A}\widehat{B}^{-1})B_1 - B_1^T\widehat{B}^{-T}A_1 + 1 \end{bmatrix}.$$

Then $B \cdot A = B_A^{-T}AB_A^{-1} \in \mathcal{A}$ implies that the lower-left entry $-B_1^T(\widehat{B}^{-T}\widehat{A}\widehat{B}^{-1})$ is 0, and since $\widehat{B}^{-T}\widehat{A}\widehat{B}^{-1}$ is invertible we must have $B_1 = 0$. Conversely, suppose $B_1 = 0$; then

$$B^{-1} = \begin{bmatrix} \widehat{B}^{-1} & 0 \\ -B_2\widehat{B} & 1 \end{bmatrix} \quad \text{and} \quad B^{-T} = \begin{bmatrix} \widehat{B}^{-T} & -\widehat{B}^{T}B_2^T \\ 0 & 1 \end{bmatrix},$$

and it follows that

$$B_A^{-T}AB_A^{-1} = \begin{bmatrix} \widehat{B}^{-T}\widehat{A}\widehat{B}^{-1} - \widehat{B}^{-T}A_1B_2\widehat{B} + \widehat{B}^{T}B_2^TB_2\widehat{B} & \widehat{B}^{-T}A_1 - \widehat{B}^{T}B_2^T \\ -B_2\widehat{B} & 1 \end{bmatrix}.$$

Then $B \cdot A = B_A^{-T}AB_A^{-1} \in \mathcal{A}$ implies that the lower-left entry $-B_2\widehat{B}$ is 0, and since $\widehat{B}$ is invertible we must have $B_2 = 0$. □

We start with a technical conjecture and lemma, that are at the heart of the proof of Proposition 44. First, recall Definition 14 of $P_{[n]}(\mathbf{f})$ for $\mathbf{f} \in (k[\mathbf{X},\mathbf{Y}])^n$:

$$P_{[n]}(\mathbf{f}) = \sum_{i=1}^{n} f_i^2 + \sum_{1\le i<j\le n} X_{ij} f_i f_j \,.$$

Conjecture 42. *The only solution* $\mathbf{f} = (f_1,\dots,f_n) \in (k[\mathbf{X},\mathbf{Y}])^n$ *to* $P_{[n]}(\mathbf{f}) = 0$ *is* $f_1 = \dots = f_n = 0$.

Remark. Suppose $n = 2$ and let $\mathbf{f} = (f_1, f_2) \in (k[\mathbf{X},\mathbf{Y}])^2$ such that $P_{[2]}(\mathbf{f}) = f_1^2 + f_2^2 + X_{12} f_1 f_2 = 0$. We can assume that f_1 and f_2 have no common factors. Rewrite the equation as $(f_1 + f_2)^2 = (2 - X_{12}) f_1 f_2$. By unique factorization in the polynomial ring $k[\mathbf{X},\mathbf{Y}]$, the irreducible polynomial $2 - X_{12}$ divides $f_1 + f_2$. Then $(2 - X_{12})^2$ divides both sides, so $2 - X_{12}$ divides, WLOG, f_1; then $2 - X_{12}$ divides f_2 as well, contradiction.

Lemma 43. *Conjecture 42 is true if we add the assumption that the* f_i *take rationals to rationals.*

Remark. We can make this assumption since k, being algebraically closed by assumption, contains a copy of $\mathbb{Q}$.

Proof. Let $\mathcal{A}_{\mathbb{Q}}$ be the set of elements of $\mathcal{A}$ whose entries are all rational. For every $A \in \mathcal{A}_{\mathbb{Q}}$, we define a bilinear form on $\mathbb{R}^n$ whose matrix M is given by $M_{ij} = 1$ if $i = j$ and $M_{ij} = M_{ji} = A_{ij}/2$ if $i < j$. We prove that if $|A_{ij}| < 1$ then M is positive definite. Let $\mathbf{v} = (v_1,\dots,v_n)^T \in \mathbb{R}^n$. Then $M = \frac{A+A^T}{2}$ so

$$2\mathbf{v}^T M\mathbf{v} = \mathbf{v}^T A\mathbf{v} + \mathbf{v}^T A^T\mathbf{v} = \sum_{i=1}^{n} 2v_i^2 + \sum_{1\le i<j\le n} 2A_{ij} v_i v_j = \sum_{1\le i<j\le n} \left(v_i^2 + 2A_{ij} v_i v_j + v_j^2\right).$$

We have two cases:

- If $v_i v_j \ge 0$, we have $v_i^2 + 2A_{ij} v_i v_j + v_j^2 = (v_i - v_j)^2 + (2A_{ij} + 2) v_i v_j \ge 0$.
- If $v_i v_j \le 0$, we have $v_i^2 + 2A_{ij} v_i v_j + v_j^2 = (v_i + v_j)^2 + (2A_{ij} - 2) v_i v_j \ge 0$.

Thus, we have $\mathbf{v}^T M\mathbf{v} \ge 0$ for all $\mathbf{v}$, which implies M is positive semidefinite. Suppose that $\mathbf{v}^T M\mathbf{v} = 0$; fix some pair $i < j$. If $v_i v_j \ge 0$, then $v_i = v_j$ and $2A_{ij} + 2 \ne 0$ so $v_i v_j = 0$ and $v_i = v_j = 0$. If $v_i v_j \le 0$, then $v_i = -v_j$ and $2A_{ij} - 2 \ne 0$ so $v_i v_j = 0$ and $v_i = v_j = 0$. Hence $\mathbf{v} = 0$. Thus M is positive definite.

Suppose that $\mathbf{X} \cup \mathbf{Y} \subset \mathbb{Q}$ and $|X_{ij}| < 1$; then each $f_i(\mathbf{X},\mathbf{Y})$ is a rational number. Letting $\mathbf{v} = (f_1(\mathbf{X},\mathbf{Y}),\dots,f_n(\mathbf{X},\mathbf{Y}))$, we have $P_{[n]}(\mathbf{f}) = \mathbf{v}^T(\frac{X+X^T}{2})\mathbf{v} = 0$ and $\frac{X+X^T}{2}$ is positive definite by the above argument, so $\mathbf{v} = (0,\dots,0)$ and $f_1(\mathbf{X},\mathbf{Y}) = 0$. This implies that the $f_i \in k[\mathbf{X},\mathbf{Y}]$ are zero within the set $\{(\mathbf{X},\mathbf{Y}) \in k^{(n^2-n)/2+m} : \mathbf{X}\cup\mathbf{Y} \subset \mathbb{Q},\ |X_{ij}| < 1\}$, thus identically zero. □

Proposition 44. *Assume Conjecture 42 and suppose that* $B \in \mathcal{B}(n)$ *is of the form* $B = \begin{bmatrix} \widehat{B} & \\ & 1 \end{bmatrix}$ *where* $\widehat{B}$ *is an* $(n-1)\times(n-1)$ *matrix. Then* $\widehat{B} \in \mathcal{B}(n-1)$.

Proof. Let $\widehat{\mathbf{X}}$ be the set of indeterminates $\{X_{ij} : 1 \le i < j \le n-1\}$. We first prove that $\widehat{B} \in M_{n-1}(\widehat{\mathbf{X}}, k)$. Since $\det B = \pm 1$, we have $\det \widehat{B} = \pm 1$. Let $b_{ij} \in k[\mathbf{X}]$ be the (i,j)th entry of $\widehat{B}^{-1}$. Then $\widehat{B}^{-T}\widehat{X}\widehat{B}^{-1} \in \widehat{\mathcal{A}}$ implies that for all ℓ we have

$$\sum_{i=1}^{n-1} b_{i\ell}^2 + \sum_{1\le i<j\le n-1} b_{i\ell} X_{ij} b_{j\ell} = (\widehat{B}^{-T}\widehat{X}\widehat{B}^{-1})_{\ell,\ell} = 1$$

where $b_{i\ell} \in k[\mathbf{X}]$. Fix $1 \leq p < n$ and let $g_i \in k[\mathbf{X} - X_{pn}]$ be the leading coefficient of $b_{i\ell}$ when considered as a polynomial in X_{pn}. Let d_i be the degree of $b_{i\ell}$ when considered as a polynomial in X_{pn}; assume that $d_1 \leq d_2 \leq \ldots \leq d_{n-1}$. Suppose for the sake of contradiction that $1 \leq d_{n-1}$. Let $n_0 > 1$ be the smallest index such that $d_{n_0} = d_{n_0+1} = \ldots = d_{n-1}$. Then we have

$$\sum_{i=n_0}^{n-1} g_i^2 + \sum_{n_0 \leq i < j \leq n-1} X_{ij} g_i g_j = 0$$

whose only solution is $g_i = 0$ for all i, by Conjecture 42; a contradiction. This shows that $\widehat{B} \in M_{n-1}(\widehat{\mathbf{X}}, k)$. Since

$$B_A^{-T} A B_A^{-1} = \begin{bmatrix} \widehat{B}^{-T} \widehat{A} \widehat{B}^{-1} & \widehat{B}^{-T} A_1 \\ & 1 \end{bmatrix} \in \mathcal{A} \text{ for all } A \in \mathcal{A}\,,$$

we have $\widehat{B}_{\widehat{A}}^{-T} \widehat{A} \widehat{B}_{\widehat{A}}^{-1} \in \widehat{\mathcal{A}}$ for all $\widehat{A} \in \widehat{\mathcal{A}}$.

Let $C \in \mathcal{B}(n)$ such that $B * C = C * B = I$. Since $C_X = B_{C \cdot X}^{-1}$, C also has the form $C = \begin{bmatrix} \widehat{C} & \\ & 1 \end{bmatrix}$ where $\widehat{C} \in M_{n-1}(\widehat{\mathbf{X}}, k)$ and $\widehat{C}_{\widehat{A}}^{-T} \widehat{A} \widehat{C}_{\widehat{A}}^{-1} \in \widehat{\mathcal{A}}$ for all $\widehat{A} \in \widehat{\mathcal{A}}$ by the above argument. Additionally, $B_{C \cdot X} C_X = I$ implies $\widehat{B}_{\widehat{C} \cdot \widehat{X}} \widehat{C}_{\widehat{X}} = I_{n-1}$; similarly $\widehat{C}_{\widehat{B} \cdot \widehat{X}} \widehat{B}_{\widehat{X}} = I_{n-1}$. Thus $\widehat{B} * \widehat{C} = \widehat{C} * \widehat{B} = I_{n-1}$ and $\widehat{B} \in \mathcal{B}(n-1)$. □

References

[1] A. I. Bondal: A symplectic groupoid of triangular bilinear forms, *Izv. RAN. Ser. Mat.*, **68**:4 (2004), 659–708.

[2] D. S. Dummit, R. M. Foote: *Abstract Algebra*, 3rd ed, Wiley & Sons Inc, p. 475.

[3] W. Magnus: Braid groups: A survey, *Proceedings of the Second International Conference on The Theory of Groups*, (1974), 463-487.

[4] David Speyer: (`mathoverflow.net/users/297`), *Commuting matrices in GL(n,Z)*, `http://mathoverflow.net/questions/55646` (version: 2011-02-16)

STUDENT ARTICLE

2 Modular Forms for the Congruence Subgroup $\Gamma(2)$

Greg Yang[†]
Harvard University '14
Cambridge, MA 02138
gyang@college.harvard.edu

Abstract

We shall find the generators of $\Gamma(2)$ and its fundamental domain. We then introduce modular functions and forms, and show that, for any modular function f, the sum of the order of f at each point in its fundamental domain and each cusp point equals half of its weight. We then characterize the vector spaces of $\Gamma(2)$ modular forms in terms of theta functions, in the process deriving the Poisson summation formula and the Jacobi identity.

2.1 The Modular Group and the Congruence Subgroups

Definition 1. The **modular group**, denoted by Γ, is the group of fractional linear transformations

$$\mathbb{C} \to \mathbb{C},\ z \mapsto \frac{az+b}{cz+d}$$

where a, b, c, d are integers such that $ad - bc = 1$.

It follows from the definition that $\Gamma \cong SL_2(\mathbb{Z})/\{I, -I\}$ by representing each element of Γ as a matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

since the composition of 2 elements of Γ is equivalent to matrix multiplication. We quotient out by $\{\pm I\}$ because the action of $-I$ is trivial. Note that $\Gamma(2)$ acts faithfully on the upper half plane.

Definition 2. A **congruence subgroup** is a subgroup of the modular group subject to certain congruence relations. In particular, the **principal congruence subgroup** mod N of Γ, denoted by $\Gamma(N)$, is the subgroup with $a \equiv d \equiv 1,\ b \equiv c \equiv 0 \mod N$. In other words, $\Gamma(N) = \{d : g \in \Gamma,\ g \equiv I \mod N\}$.

In the most basic case, we have $\Gamma(1) = \Gamma$. In this article we deal with $\Gamma(2)$. For the discussion of Γ itself and its modular forms, see [1, 2].

Let $\mathbb{H}$ be the upper half plane, i.e., the set $\{z \in \mathbb{C} : \mathrm{Re}(z) > 0\}$. First we find the **fundamental domain** of $\Gamma(2)$. This is the region R in $\mathbb{H}$ such that for every $z \in \mathbb{H}$, $R \cap \Gamma z$ is a singleton. In fact we shall prove the following theorem:

Theorem 3. *Let D be the region in $\mathbb{H}$ bounded by the lines $\mathrm{Re}(z) = \pm 1$ and the semicircles $|z \pm 1/2| = 1/2$. Then*

[†]Greg Yang '14 concentrates in mathematics at Harvard University. He enjoys tackling difficult problems in mathematics and computer science. In addition, he is a drummer, an electronic musician, and dubstep artist.

1. *If X is the subgroup of $\Gamma(2)$ generated by $S_2 = z/(2z+1)$ and $T_2 = z+2$, then for every $z \in \mathbb{C}$ there exists a $U \in X$ such that $Uz \in D$. Furthermore if z is also in D then $U = I$. Hence, D is the fundamental domain of X.*

2. *For every $z \in \mathbb{C}$ there exists a $U \in \Gamma(2)$ such that $Uz \in D$. Furthermore if z is also in D then $U = I$. Hence, D is the fundamental domain of $\Gamma(2)$.*

3. *X is actually equal to $\Gamma(2)$. Equivalently, S_2 and T_2 generate $\Gamma(2)$.*

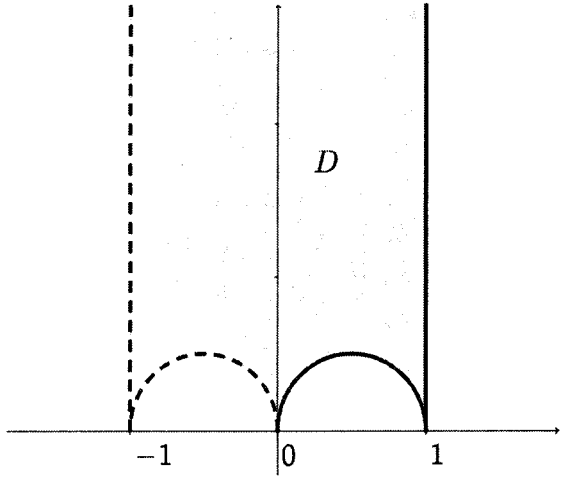


Figure 2.1: The fundamental domain of $\Gamma(2)$

First, we prove a simple lemma.

Lemma 4. *If* $U = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ *with* $U \in \Gamma$, *then* $\operatorname{Im}(Uz) = \frac{\operatorname{Im}(z)}{|cz+d|^2}$

The proof is a straightforward calculation:

$$\begin{aligned} Uz &= \frac{az+b}{cz+d} \\ &= \frac{(az+b)(c\overline{z}+d)}{|cz+d|^2} \\ &= \frac{ac|z|^2 + bd + adz + bc\overline{z}}{|cz+d|^2} \\ &= \frac{ac|z|^2 + bd + ad(z+\overline{z}) + (bc-ad)\overline{z}}{|cz+d|^2} \end{aligned}$$

As $ac|z|^2 + bd + ad(z+\overline{z})$ is real, and $(bc-ad)\overline{z} = -\overline{z}$ (by the SL condition) has imaginary part $\operatorname{Im}(z)$, we get the desired result.

Now, for any $z \in \mathbb{C}$, $\operatorname{Im}(Uz) = \frac{\operatorname{Im}(z)}{|cz+d|^2}$ has a maximum because $|cz+d| \geq |c| \operatorname{Im}(z) \geq \operatorname{Im}(z)$. Choose $U' \in X$ that maximizes $\operatorname{Im}(U'z)$. Let n be the integer such that $|\operatorname{Re}(T_2^n U'z)| \leq 1$. Let Θ be the open disk centered at $-1/2$ with radius 1/2. If $z' = T_2^n U'z$ falls in Θ, then

$$|2z'+1|^{-1} > 1 \Rightarrow \operatorname{Im}(S_2 z') = \operatorname{Im}(z')|2z'+1|^{-1} > \operatorname{Im}(z')$$

which contradicts the optimality of U'. Thus, z' is outside Θ. A similar argument with $S_2\Theta$ and the operator S_2^{-1} shows that z' is outside $S_2\Theta$. Therefore z' must be in the fundamental domain or on its border. If z' is on the left border, then apply T_2 or S_2 to move it to the right border. This shows that D has at least one representative of each $z \in \mathbb{C}$ under the action of X. If we switch X with $\Gamma(2)$, the proof gives the same result.

Now we show that each of these representatives is the unique one in its equivalence class. Consider the images of the unit disk under the inverse maps of $z \mapsto 2nz + (2m+1)$ for $n \neq 0$: They are disks centered at $-\frac{2m+1}{2n}$ with radius $|1/2n|$. Hence they never intersect D. For that reason, $|2nz + (2m+1)| > 1$ and $\mathrm{Im}(Uz) = \frac{\mathrm{Im}(z)}{|cz+d|^2} < \mathrm{Im}(z)$ if $z \in D$. Furthermore,

$$U = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & b \\ 2n & 2m+1 \end{pmatrix}.$$

But U has an inverse, so this inequality cannot hold for U^{-1} at the same time. Hence $n = c = 0$.

Even if $c = 0$, $\mathrm{Im}(Uz) = \mathrm{Im}(z)$ only if $d = \pm 1 = a$. Thus 2 representatives of the same class must be related by some power of T_2 if they are both in D. But this is clearly impossible. This proof again works for both X and $\Gamma(2)$.

Thus we have proven points 2 and 1. These in turn show point 3, as each $U \in \Gamma(2)$ that carries w to $z \in D$ must be unique (as the inverse is unique), and hence is equal to some element of X.

2.2 Modular Functions

Definition 5. For an integer k, define the weight-$2k$ action of Γ (or any of its subgroups) on the functions $f : \mathbb{H} \to \mathbb{H}$ thus: For $U = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$,

$$f[U](z) = (cz+d)^{-2k} f\left(\frac{az+b}{cz+d}\right).$$

It is not hard to check that $f[U][V] = f[UV]$ so this action is well-defined.

Definition 6. A function $f : \mathbb{H} \to \mathbb{H}$ is a **weakly modular function** of weight $2k$ with respect to a congruence subgroup Γ' if it is meromorphic and $f[U] = f$ for all $U \in \Gamma'$ (where $[U]$ is the action of weight $2k$).

In the case of $\Gamma' = \Gamma(2)$, the second condition in the definition ($f[U] = f$ for all $U \in \Gamma'$) is equivalent to

$$f(z) = f(z+2) = (2z+1)^{-2k} f\left(\frac{z}{2z+1}\right)$$

by Theorem 3.

Let Δ be the open unit disk and $\Delta^* := \Delta - \{0\}$. The 2-periodicity implies the existence of a meromorphic function $\tilde{f} : \Delta^* \to \mathbb{C}$ such that $\tilde{f}(e^{i\pi z}) = f(z)$.

Note further that as $U\Gamma(2)U^{-1} \equiv I \mod 2$, $\Gamma(2)$ is normal. Hence, for any $U \in \Gamma$ and $W \in \Gamma(2)$ there exists $V \in \Gamma(2)$ such that $VU = UW$ and therefore $f[V][U] = f[U][W]$. Thus we have shown

Proposition 7. *If f is a weakly modular function of $\Gamma(2)$, then so is $f[U]$ for any U.*

We may now define

$$\tilde{f}'(e^{i\pi z}) = f[-1/w](z) \text{ and } \tilde{f}''(e^{i\pi z}) = f[1-1/w](z)$$

in the same manner as for $\tilde{f}$.

Definition 8. A weakly modular function f with respect to $\Gamma(2)$ is said to be a **modular function** if $\tilde{f}, \tilde{f}', \tilde{f}''$ all extend to meromorphic functions in Δ. Equivalently, we require them to be meromorphic at 0.

Let $v_p(f)$ represent the order of a meromorphic function f at point p, i.e. if f has an expansion $a_k(z-p)^k + a_{k+1}(z-p)^{k+1} + \cdots$ where $a_k \neq 0$, then $v_p(f) = k$. If f is a modular function with respect to $\Gamma(2)$, then also define $v_\infty(f) = v_0(\tilde{f})$, $v_0(f) = v_0(\tilde{f}')$, and $v_1(f) = v_0(\tilde{f}'')$.

Theorem 9. *If f is a modular function of weight $2k$ with respect to $\Gamma(2)$, then*

$$v_\infty(f) + v_0(f) + v_1(f) + \sum_{p \in D} v_p(f) = k$$

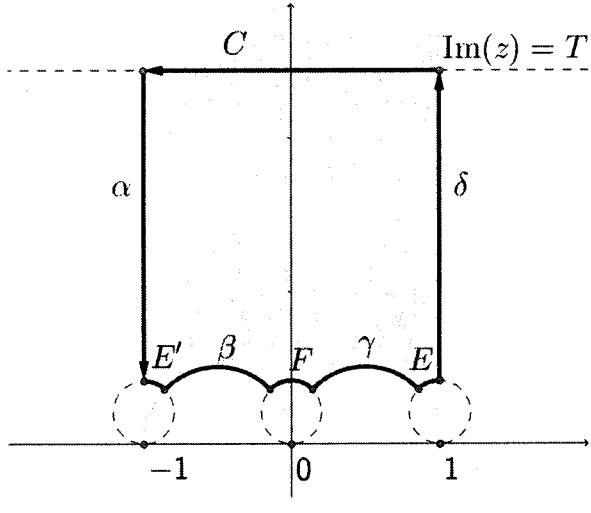


Figure 2.2: The integration path

For a real positive number T, denote the segment from $1 + Ti$ to $-1 + Ti$ as C, and let F be the image of C under $z \mapsto -\frac{1}{z}$. Let E be the left half of the image of C under the transformation $z \mapsto 1 - \frac{1}{z}$, and E' be the right half of it but translated 2 units to the left so that it cuts D. Let $\alpha, \beta, \gamma, \delta$ be the rest of the boundaries of D cut by C, F, E, and E', as shown in Figure 2.2.

Since $\tilde{f}$ is meromorphic on the unit disc, for some punctured disc around 0, $\tilde{f}$ cannot have any root or pole. This translates to the statement that f has no root or pole for $\mathrm{Im}(z) > t$ for some t. Similarly, by applying the same reasoning to $\tilde{f}'$ ($\tilde{f}''$), we see that in some punctured half disk around 0, 1 or -1 (equivalent under the action of $\Gamma(2)$) no root or pole can be present.

Consequently, for a large enough T, the region within $b = C + \alpha + E' + \beta + F + \gamma + E + \delta$ contain all the poles and roots in D.

Recall the *argument principle* [1]: For a meromorphic function f, the number of roots minus the number of poles, counting multiplicity, inside a region R is given by:

$$\frac{1}{2\pi i} \int_{\partial R} \frac{f'}{f}\, dz = \frac{1}{2\pi i} \int_{\partial R} d(\log f)$$

if no root or pole lies on ∂R, the boundary of R.

We may first assume no roots or poles lie on B. By the above principle we have $\sum_{p\in D} v_p(f) = \frac{1}{2\pi i}\int_B d(\log f)$. We break down the evaluation of this integral to the following steps.

1. Immediately, the integral $\int_\alpha + \int_\delta$ is 0 as $f(z+2) = f(z)$.
2. Since S_2 maps β to $-\gamma$, we have

$$\begin{aligned}\int_\beta + \int_\gamma d(\log f) &= \int_\beta d(\log f(z)) - d(\log f)\left(\frac{z}{2z+1}\right)\\ &= \int_\beta d(\log f(z)) - d(\log(2z+1)^{2k} f(z))\\ &= -2k\int_\beta d(\log(2z+1))\end{aligned}$$

   As $T \to \infty$, this integral approaches $-2k(-\pi i) = 2\pi i k$.
3. Because $z \mapsto \exp(i\pi z)$ maps C to a negatively-oriented circle ω around 0, $\int_C d(\log f(z)) = \int_\omega d(\log \tilde{f}(z)) = -2\pi i v_\infty(f)$.
4. Similarly,

$$\begin{aligned}\int_F d(\log f(z)) &= \int_C d(\log f(-1/z))\\ &= \int_C d(\log \tilde{f}'(e^{i\pi z}) + 2k\log z)\\ &= \int_\omega d(\log \tilde{f}'(z)) + 2k\int_C d(\log z)\\ &= -2\pi i v_0(f) + 2k\int_C d(\log z)\end{aligned}$$

   As T goes to infinity, $\int_C d(\log z)$ goes to 0, so the integral is just $-2\pi i v_0(f)$.

   Repeating the process in 1 by replacing F with $E + E'$, we find $\int_{E+E'} d(\log f(z))$ goes to $-2\pi i v_1(f)$ as $T \to \infty$.

Putting these results together, we get

$$\sum_{p\in D} v_p(f) = \frac{1}{2\pi i}\int_B d(\log f) = k - (v_\infty(f) + v_0(f) + v_1(f))$$

as $T \to \infty$. Since this equality is independent of T, we arrive at our desired result.

Now what if there were a pole or root at the point p on the contour α? Then by the modular relation, there must also be one at $p+2$ on δ. In this case, we integrate along a detour around the zero or pole. Our choice of path will be a small semicircle ϵ around p going out of D on α and going into D on δ, to maintain symmetry with respect to the modular group, and chosen so that no pole or root lies on ϵ, as shown in Figure 2.3. Note that we may always select such a contour since the zeros and poles of a meromorphic function are isolated.

We may also have a root or pole at the point q on β. Again, by modularity, then there must be one at $q/(2q+1)$ on γ. Again, draw a small arc ϵ' centered at q going out of D and the corresponding arc $S_2\epsilon'$ into D, chosen so neither have any roots or poles lying on them. The new β' and γ' still satisfy $S_2\beta' = -\gamma'$.

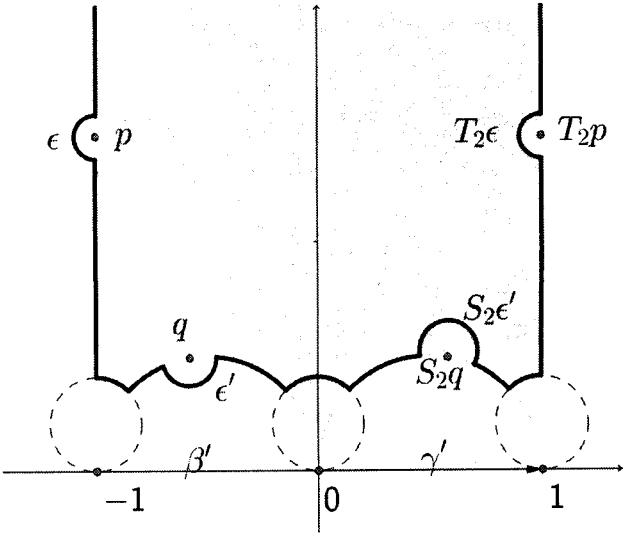


Figure 2.3: Detour around poles or roots

2.3 Modular Forms

Definition 10. A modular function f is a **modular form** (with respect to $\Gamma(2)$) if it is holomorphic, and $\tilde{f}$, $\tilde{f}'$, and $\tilde{f}''$ extend to holomorphic functions on Δ.

Definition 11. If f is a $\Gamma(2)$ modular forms, the points $0, 1, \infty$ are called the **cusp points**, and $\tilde{f}(0)$, $\tilde{f}'(0)$, and $\tilde{f}''(0)$ are called the **cusp values**. A modular form is called a **cusp form** if the cusp values are all 0.

Let's first look at a few examples of modular forms of $\Gamma(2)$.

2.3.1 Theta Functions

For the purposes of this article (see [3] for more about the more general theta function of 2 complex variables), define

$$\begin{aligned}\vartheta_{00}(z) &= \sum_{n\in\mathbb{Z}} e^{i\pi n^2 z},\\ \vartheta_{01}(z) &= \sum_{n\in\mathbb{Z}} (-1)^n e^{i\pi n^2 z},\\ \vartheta_{10}(z) &= \sum_{n\in\mathbb{Z}} e^{i\pi (n+1/2)^2 z}.\end{aligned}$$

All 3 sums are absolutely convergent on any half plane $\mathrm{Im}(z) > t > 0$, as the terms decrease in magnitude at a rate of $e^{-n^2\,\mathrm{Im}(z)}$. Hence, they define holomorphic functions on $\mathbb{H}$.

Immediately from their equations, we obtain

1. $\vartheta_{00}(z+1) = \vartheta_{01}(z)$, which implies that ϑ_{00} and ϑ_{01} are 2-periodic.

2. $\vartheta_{10}(z+1) = \sum_{\mathbb{Z}} e^{i\pi(n+1/2)^2 z} e^{i\pi(n+1/2)^2}$. As $(2n+1)^2$ is 1 mod 8, this is just $e^{i\pi/4}\vartheta_{10}(z)$.

Our goal is to show that $\vartheta_{00}^4, \vartheta_{01}^4, \vartheta_{10}^4$ are modular forms, and to do this we need to find out their behavior under the action of S_2. We need some intermediary tools first.

Theorem 12. *(Poisson Summation Formula) If $f : \mathbb{R} \to \mathbb{C}$ is such that $\sum_{n\in\mathbb{Z}} f(n)$ and $\sum_{n\in\mathbb{Z}} \hat{f}(n)$ are absolutely convergent, then*

$$\sum_{n\in\mathbb{Z}} f(n) = \sum_{n\in\mathbb{Z}} \hat{f}(n)$$

where $\hat{f}$ is the Fourier transform of f.

Let $F(x) = \sum_{n\in\mathbb{Z}} f(x+n)$. F has period 1 and thus has a Fourier expansion. Its kth coefficient is

$$\begin{aligned}
& \int_0^1 F(x)e^{-2ki\pi x}dx \\
= & \int_0^1 \sum_{n\in\mathbb{Z}} f(x+n)e^{-2ki\pi x}dx \\
= & \sum_{n\in\mathbb{Z}} \int_0^1 f(x+n)e^{-2ki\pi x}dx \qquad \text{by absolute convergence} \\
= & \int_{-\infty}^{\infty} f(x)e^{-2ki\pi x}dx \\
= & \hat{f}(k)
\end{aligned}$$

Thus we can write

$$F(x) = \sum_{n\in\mathbb{Z}} \hat{f}(n)e^{2ni\pi x}$$

Hence $F(0) = \sum_{n\in\mathbb{Z}} \hat{f}(n)$ as desired.

Corollary 13. *$\vartheta_{00}(-1/z) = \sqrt{z/i}\vartheta_{00}(z)$, where $\sqrt{z}$ denotes the branch giving the value with nonnegative imaginary part.*

Note that the left and the right sides are both holomorphic functions on $\mathbb{H}$, and by uniqueness of analytic conitnuation we only need to show this for $z = it^2$ for $t > 0$. So our claim is

$$\sum_{n\in\mathbb{Z}} e^{-\pi(n/t)^2} = t\sum_{n\in\mathbb{Z}} e^{-\pi(nt)^2}$$

But as $e^{-\pi(x/t)^2}$ (as a function of x) is the Fourier transform of $te^{-\pi(xt)^2}$, this is the statement in this case of the Poisson Summation Formula.

Corollary 14. *With the notation above, $\vartheta_{01}(-1/z) = \sqrt{z/i}\vartheta_{10}(z)$. Equivalently, $\vartheta_{00}(1-1/z) = \sqrt{z/i}\vartheta_{10}(z)$.*

Using the fact that $e^{-\pi(x+1/2)^2t^2}$ has Fourier transform $t^{-1}e^{-\pi x^2/t^2+i\pi x}$, one applies the same reasoning as above.

Theorem 15. *$\vartheta_{00}^4, \vartheta_{01}^4, \vartheta_{10}^4$ are modular forms of weight 2 with respect to $\Gamma(2)$.*

We have already seen the 2-periodicity of ϑ_{00} and ϑ_{01}, and $\vartheta_{10}(z+1) = e^{i\pi/4}\vartheta_{10}(z)$ gives $\vartheta_{10}^4(z+1) = -\vartheta_{10}^4(z)$ and hence 2-periodicity. Corollaries 13 and 14 give (respectively)

$$\begin{aligned}
\vartheta_{00}^4[-1/z] &= -\vartheta_{00}^4 \\
\vartheta_{01}^4[-1/z] &= -\vartheta_{10}^4
\end{aligned}$$

As $(-1/z) \circ T_2 \circ (-1/z) = S_2^{-1}$, we have

$$\begin{aligned} \vartheta_{00}^4[S_2^{-1}] &= \vartheta_{00}^4[-1/z][T_2][-1/z] \\ &= -\vartheta_{00}^4[T_2][-1/z] \\ &= -\vartheta_{00}^4[-1/z] \\ &= \vartheta_{00}^4 \end{aligned}$$

So $\vartheta_{00}^4[S_2] = \vartheta_{00}^4$ as well, and hence ϑ_{00}^4 is weakly modular. By proposition 7, so is ϑ_{01}^4, and, by corollary 14, so is ϑ_{10}^4.

As

$$\begin{aligned} \tilde{\vartheta_{00}^4}'(e^{i\pi z}) &= \vartheta_{00}^4[-1/w](z) = -\vartheta_{00}^4(z) \\ \tilde{\vartheta_{00}^4}''(e^{i\pi z}) &= \vartheta_{00}^4[1-1/w](z) = \vartheta_{10}^4(z) \\ \tilde{\vartheta_{01}^4}'(e^{i\pi z}) &= \vartheta_{01}^4[-1/w](z) = -\vartheta_{10}^4(z) \\ \tilde{\vartheta_{01}^4}''(e^{i\pi z}) &= \vartheta_{00}^4[1-1/w](z) = -\vartheta_{00}^4(z) \\ \tilde{\vartheta_{10}^4}'(e^{i\pi z}) &= \vartheta_{10}^4[-1/w](z) = -\vartheta_{01}^4(z) \\ \tilde{\vartheta_{10}^4}''(e^{i\pi z}) &= \vartheta_{10}^4[1-1/w](z) = \vartheta_{01}^4(z) \end{aligned}$$

So to prove that they are modular forms, it is enough to verify that none of $\vartheta_{00}^4, \vartheta_{01}^4, \vartheta_{10}^4$ diverge as $z \to i\infty$.

Indeed, as $\vartheta_{00}(z) = \sum_{-\infty}^{\infty} e^{i\pi n^2 z}$ is uniformly convergent on any half plane $\mathrm{Im}(z) > t > 0$, we can evaluate term by term when seeking the limit as $z \to i\infty$: every term except for $n = 0$ drops out, so $\vartheta_{00}(\infty) = 1$. Similarly, $\vartheta_{01}(\infty) = 1$ and $\vartheta_{10}(\infty) = 0$. So we can tabulate

	$\tilde{f}'(0)$	$\tilde{f}''(0)$	$\tilde{f}(0)$
ϑ_{00}^4	-1	0	1
ϑ_{01}^4	0	-1	1
ϑ_{10}^4	-1	1	0

Figure 2.4: cusp values of theta functions

2.3.2 Vector Spaces of $\Gamma(2)$ Modular Forms

Definition 16. The modular forms of weight $2k$ with respect to $\Gamma(2)$ clearly form a vector space, which we will denote M_k^2. The cusp forms of weight $2k$ with respect to $\Gamma(2)$ also form a subspace (it is in fact the kernel of the map $f \mapsto (\tilde{f}(0), \tilde{f}'(0), \tilde{f}''(0)) : M_k^2 \to \mathbb{C}^3$) which we will denote N_k^2.

It is immediate that $M_0^2 \cong \mathbb{C}$. For M_1^2, we will prove

Theorem 17. *$\dim M_1^2 = 2$, and M_1^2 has basis $\{\vartheta_{01}^4, \vartheta_{10}^4\}$*

First, the image of the projection map will have at most 2 dimensions. Indeed, the sum $\tilde{f}(0) + \tilde{f}'(0) + \tilde{f}''(0)$ must equal zero in this case. Defining C as in the proof of Theorem 9 and ω as the image of C under $z \mapsto e^{i\pi z}$, we have

$$\begin{aligned} \tilde{f}(0) + \tilde{f}'(0) + \tilde{f}''(0) &= \int_\omega \frac{\tilde{f} + \tilde{f}' + \tilde{f}''}{q} dq \qquad \text{by the residue theorem [1]} \\ &= i\pi \int_C f(z)dz + f(Sz)d(Sz) + f(TSz)d(TSz) \end{aligned}$$

where $S(z) = -1/z$ and $Tz = z + 1$. Then $S^{-1}C = F$ and $(TS)^{-1}C \equiv E + E'$ as in the proof of Theorem 9. So this integral is just

$$i\pi \left(\int_C + \int_F + \int_{E+E'} \right) f(z)dz$$

Again, define B as defined in the proof of Theorem 9. Then $\int_B f dz = 0$ because $f(z)$ is holomorphic inside. The left vertical side cancels with the right ($\int_\alpha + \int_\delta = 0$), and

$$\int_\beta f(z)dz = \int_\beta f\left(\frac{z}{2z+1}\right) d\frac{z}{2z+1} = \int_{-\gamma} f(z)dz.$$

So it must be that

$$\left(\int_C + \int_F + \int_{E+E'} \right) f(z)dz = \int_{B-\alpha-\beta-\gamma-\delta} f(z)dz = 0$$

and therefore the dimension of M_1^2/N_1^2 is at most 2.

Note that $N_1^2 = 0$ as Theorem 9 implies that the cusp values cannot all be 0. As ϑ_{01}^4 and ϑ_{10}^4 are linearly independent, evident from Figure 2.4, they form a basis for M_1^2.

□

Observe that because $\vartheta_{00}^4 - \vartheta_{01}^4 - \vartheta_{10}^4$ has zeroes at all cusps but is still a weight-2 modular form, it is identically 0. Thus deduce the Jacobi Identity.

Corollary 18. *(Jacobi Identity)* $\vartheta_{00}^4 = \vartheta_{01}^4 + \vartheta_{10}^4$.

It is not hard to see that $\vartheta_{00}^8, \vartheta_{01}^8, \vartheta_{10}^8$ forms a basis of M_2^2, as there are no weight-4 cusp forms. We have thus proved that $\dim M_0^2 = 1$, $\dim M_1^2 = 2$, $\dim M_2^2 = 3$.

Now notice that Theorem 9 for $k = 1$ implies that the theta functions are nonzero on $\mathbb{H}$. Hence, $\xi := (\vartheta_{00}\vartheta_{01}\vartheta_{10})^4$ has simple zeroes at all 3 cusps and is nonzero on $\mathbb{H}$. Thus it is a cusp form (it is, in fact, the one with minimum weight). Hence there is an isomorphism between M_{k-3}^2 and N_k^2 defined by $f \mapsto \xi f$. Consequently, $\dim M_k^2 = 3 + \dim M_{k-3}^2$ and, inductively, $\dim M_k^2 = k + 1$.

Observe that $\{\vartheta_{00}^{8k}, \vartheta_{01}^{8k}, \vartheta_{10}^{8k}\}$ form a basis of M_{2k}^2/N_{2k}^2, and $\{\vartheta_{00}^{8k}\vartheta_{01}^4, \vartheta_{01}^{8k}\vartheta_{10}^4, \vartheta_{10}^{8k}\vartheta_{00}^4\}$ form a basis of M_{2k+1}^2/N_{2k+1}^2. We can therefore write the basis of M_k^2 in terms of the monomials in $\vartheta_{00}^4, \vartheta_{01}^4, \vartheta_{10}^4$. In fact, since $\vartheta_{00}^4 = \vartheta_{01}^4 + \vartheta_{10}^4$, we only need to write the basis in terms of ϑ_{01}^4 and ϑ_{10}^4.

We shall prove that the $k + 1$ monomials $\vartheta_{01}^{4b}\vartheta_{10}^{4c}$ with $b + c = k$, $b, c \geq 0$ form a basis of M_k^2. We only need linear independence as the size of such a set matches the dimension of the vector space. If there is some nontrivial linear combination of these monomials that equals 0, then $\vartheta_{01}^4/\vartheta_{10}^4$ satisfies a nontrival polynomial. But, as such a polynomial has discrete roots, $\vartheta_{01}^4/\vartheta_{10}^4$ must be constant. This is clearly false. We have thence arrived at

Theorem 19. *M_k^2 has dimension $k + 1$ and basis* $\{\vartheta_{01}^{4b}\vartheta_{10}^{4c} : b + c = k,\ b, c \geq 0\}$. ■

2.4 Ways of Extending Our Results

There are many more relations to discover between $\Gamma(2)$ and Γ modular forms. For example, if $\Lambda = g_2^3 - 27g_3^2$ is the modular discriminant (the Γ cusp form of the lowest weight), then one may find $\sqrt{\Lambda} = 1/\sqrt{2}\xi$ (for details in defining g_2 and g_3, see [1, 276]). At the same time, $\Gamma(2)$ modular forms are related intimately with elliptic curves and Weierstrass functions. The roots of the cubic $4\wp^3 - g_2\wp - g_3$ associated with the Weierstrass function $\wp(z)$ with periods 1 and τ, in particular, can be expressed thus

$$\begin{aligned} e_1(\tau) &= \frac{1}{3}\pi^2(\vartheta_{00}^4(\tau) + \vartheta_{01}^4(\tau)) \\ e_2(\tau) &= -\frac{1}{3}\pi^2(\vartheta_{00}^4(\tau) + \vartheta_{10}^4(\tau)) \\ e_3(\tau) &= \frac{1}{3}\pi^2(\vartheta_{10}^4(\tau) - \vartheta_{01}^4(\tau)) \end{aligned}$$

The theory of modular functions and forms developed in this paper has hopefully made a jumping board for the avid reader to actively explore these connections.

References

[1] L. V. Ahlfors. *Complex Analysis: An Introduction to the Theory of Analytic Functions of One Complex Variable*. McGraw Hill, Inc, 1979.

[2] Jean Pierre Serre. *A Course in Arithmetic*. Springer-Verlag, 1973.

[3] Elias M. Stein and Shakarchi. *Complex Analysis*. Princeton University Press, 2003.

STUDENT ARTICLE

3 The δ Function as a Measure

Thomas Meyer[†]
Columbia University '13
New York, NY 10027
tkm2115@columbia.edu

Abstract

The δ function was formulated by the theoretical physicist Paul A. M. Dirac in his book *The Principles of Quantum Mechanics*, first published in 1930. This function satisfies the following conditions:

$$\int_{-\infty}^{\infty} \delta(x)\,dx = 1, \quad \delta(x) = 0 \text{ for } x \neq 0.$$

Because the integral above is a Riemann integral, there can be no function that meets both of these conditions. To bypass this difficulty, Dirac defined δ to be a generalized function. It is this definition which is found in the majority of the literature on the δ function. However, this article takes a different approach. Namely, the δ function is defined as a measure, the Dirac measure. Further, it is proven using Lebesgue integration that all the usual properties of the δ function hold for the Dirac measure.

3.1 Introduction

The δ function (or the Dirac delta function) was formulated by the theoretical physicist Paul A. M. Dirac in his book *The Principles of Quantum Mechanics*, first published in 1930. This function has proven useful to many fields, among them Fourier theory, quantum mechanics, signal processing, and electrodynamics. Dirac's definition [1] of the δ function is as follows:

> *we introduce a quantity $\delta(x)$ depending on a parameter x satisfying the conditions*
>
> $$\int_{-\infty}^{\infty} \delta(x)dx = 1,$$
>
> $$\delta(x) = 0 \textit{ for } x \neq 0.$$

However, since any real function which is nonzero only at one point must vanish when integrated from $-\infty$ to ∞, there can be no function which satisfies both conditions. To bypass this difficulty, many textbooks state that δ is a generalized function. Further, because the notion of a generalized function requires a familiarity with functional analysis, the more introductory of these textbooks do not provide proofs of the δ function's properties. Instead, they appeal to intuitive, albeit incorrect, arguments. For instance, Dirac gives the following argument for the statement $\int_{-\infty}^{\infty} \delta(x)\,dx = 1$:

> *To get a picture of $\delta(x)$, take a function of the real variable x which vanishes everywhere except inside a small domain, of length ϵ say, surrounding the origin $x = 0$, and which is so large inside this domain that its integral over this domain is unity. The exact shape of the function inside this domain does not matter, provided there are no unnecessarily wild variations... Then in the limit $\epsilon \to 0$, this function will go over into $\delta(x)$. [1]*

[†]Thomas Meyer is a second semester junior at Columbia University's School of Engineering and Applied Science, where he is majoring in applied mathematics. He transferred there from Bard College at Simon's Rock, where he majored in pure mathematics and computer science.

There is an alternative to this characterization of the δ function as a generalized function: defining the δ function as a measure. It is the purpose of the remainder of this article to define the δ function in this way and to rigorously prove, using Lebesgue integration, that all the properties attributed to the δ function also hold for this measure.

3.2 Defining the Measure

Definition 1. Let $\mathcal{A}$ be the power set of $\mathbb{R}$, and fix $c \in \mathbb{R}$. Consider a function $\delta_c : \mathcal{A} \to \{0, 1\}$ defined as follows:

$$\delta_c(A) = \begin{cases} 0 : c \notin A \\ 1 : c \in A \end{cases} \quad [2, 3].$$

Theorem 2. *The function δ_c is a measure, commonly referred to as the "Dirac measure."*

Proof. The domain of δ_c, $\mathcal{A} = \mathcal{P}(\mathbb{R})$, is a well-known σ-algebra. By definition, $\delta_c(A) \geq 0$ for any $A \in \mathcal{A}$. Further, because $c \notin \emptyset$, $\delta_c(\emptyset) = 0$. We next need to verify that

$$\delta_c\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \delta_c(E_i)$$

where $\{E_i\}_{i=1}^{\infty}$ is a countable collection of pairwise disjoint sets in $\mathcal{A}$. Two cases must be considered:

$$c \notin \bigcup_{i=1}^{\infty} E_i, \tag{3.1}$$

$$c \in \bigcup_{i=1}^{\infty} E_i. \tag{3.2}$$

In case (1), none of the E_i contain c. Therefore, we have

$$\delta_c\left(\bigcup_{i=1}^{\infty} E_i\right) = 0 = \sum_{i=1}^{\infty} 0 = \sum_{i=1}^{\infty} \delta_c(E_i).$$

Due to the fact that the E_i are disjoint, in case (2), there exists exactly one $E_k \in \{E_i\}_{i=1}^{\infty}$ such that $c \in E_k$. That is, $\delta_c(E_k) = 1$ and $\delta_c(E_i) = 0$ for all $i \neq k$. Therefore, we have

$$\delta_c\left(\bigcup_{i=1}^{\infty} E_i\right) = 1 = \delta_c(E_k) = \sum_{i=1}^{\infty} \delta_c(E_i).$$

Having established the countable additivity of δ_c in cases (1) and (2), δ_c must be countably additive. □

Example 3. Take the function

$$\hat{\delta}(x) = \sum_{c=-\infty}^{\infty} \delta_c(\{x\}),$$

where $x \in \mathbb{R}$. This function can be visualized as a sequence of unit impulses, where each impulse is centered at an integer. In this way, $\hat{\delta}$ corresponds to the Dirac comb, which is given by:

$$\Delta(x) = \sum_{n=-\infty}^{\infty} \delta(x - n).$$

The Dirac comb is often used in digital signal processing. Specifically, it is used in the mathematical modeling of the reconstruction of a continuous signal from equally spaced samples.

3.3 Properties of δ_c

Lemma 4. *Suppose that g is a nonnegative, simple, $\mathcal{A}$-measurable function defined on $\mathbb{R}$. Then the Dirac measure satisfies the equation*

$$\int_{\mathbb{R}} g\, d\delta_c = g(c).$$

Proof. Take $(a_k)_{k=1}^n$, for some $n \in \mathbb{Z}^+$, to be the n distinct values of g. Additionally, let $a_l = g(c)$, where $1 \leq l \leq n$. Denote by $(A_k)_{k=1}^n$ the n sets such that $A_k = \{x \in \mathbb{R} : g(x) = a_k\}$. Clearly, $\delta_c(A_l) = 1$ and $\delta_c(A_k) = 0$ for any $k \neq l$. Thus,

$$\int_{\mathbb{R}} g\, d\delta_c = \sum_{k=1}^{n} a_k \delta_c(A_k) = a_l = g(c).$$

□

Corollary 5. *The Dirac measure has the following property:*

$$\int_{\mathbb{R}} d\delta_c = 1.$$

This corollary formalizes the equality $\int_{-\infty}^{\infty} \delta(x)\, dx = 1$.

Proof. The function $f = 1$ is trivially simple and nonnegative. Further, f is $\mathcal{A}$-measurable. To demonstrate this, first select some $t \in \mathbb{R}$ satisfying $t > 1$. Then $\{x : f(x) < t\} = \mathbb{R} \in \mathcal{A}$. If t is chosen so that $t \leq 1$, then $\{x : f(x) < t\} = \emptyset \in \mathcal{A}$. With the $\mathcal{A}$-measurability of f proved, the previous lemma can be applied. This gives

$$\int_{\mathbb{R}} d\delta_c = \int_{\mathbb{R}} f\, d\delta_c = f(c) = 1.$$

□

Theorem 6. *Assume that f is an $\mathcal{A}$-measurable function such that $|f(c)| < \infty$. The Dirac measure then gives:*

$$\int_{\mathbb{R}} f\, d\delta_c = f(c).$$

This theorem corresponds to the most important property of the δ-function, the equation

$$\int_{-\infty}^{\infty} f(x)\delta(x - c)\, dx = f(c).$$

The theorem states that when δ_c is the measure associated with the integral, then provided that f meets certain conditions, the value of f at c is "picked out."

Proof. The function f is integrable over $\mathbb{R}$ if $\int_{\mathbb{R}} f^+ d\delta_c < \infty$ and $\int_{\mathbb{R}} f^- d\delta_c < \infty$, where:

$$f^+ = \begin{cases} f(x) & f(x) > 0, \\ 0 & \text{otherwise} \end{cases}$$

and

$$f^- = \begin{cases} -f(x) & f(x) < 0, \\ 0 & \text{otherwise.} \end{cases}$$

Because f^+ is a nonnegative, $\mathcal{A}$-measurable function, there exists a sequence of nonnegative, simple, $\mathcal{A}$-measurable functions, $(f_n)_{n=1}^{\infty}$, which satisfies the following two conditions:

$$0 \leq f_n(x) \leq f_{n+1}(x) \text{ for all } n \in \mathbb{N} \text{ and } x \in \mathbb{R}$$

and

$$\lim_{n\to\infty} f_n(x) = f^+(x) \text{ for all } x \in \mathbb{R}.$$

Likewise, since f^- is also a nonnegative, $\mathcal{A}$- measurable function, there exists a sequence of nonnegative, simple, $\mathcal{A}$-measurable functions, $(g_n)_{n=1}^{\infty}$, that fulfills:

$$0 \leq g_n(x) \leq g_{n+1}(x) \text{ for all } n \in \mathbb{N} \text{ and } x \in \mathbb{R}$$

and

$$\lim_{n\to\infty} g_n(x) = f^-(x) \text{ for all } x \in \mathbb{R}.$$

By the previous lemma, we have

$$\int_{\mathbb{R}} f_n \, d\delta_c = f_n(c) \text{ and } \int_{\mathbb{R}} g_n \, d\delta_c = g_n(c).$$

Therefore, by the Lebesgue Monotone Convergence Theorem (see [4]),

$$\int_{\mathbb{R}} f^+ d\delta_c = \lim_{n\to\infty} \int_{\mathbb{R}} f_n \, d\delta_c = \lim_{n\to\infty} f_n(c) = f^+(c).$$

Applying this same theorem again gives

$$\int_{\mathbb{R}} f^- d\delta_c = \lim_{n\to\infty} \int_{\mathbb{R}} g_n \, d\delta_c = \lim_{n\to\infty} g_n(c) = f^-(c).$$

Moreover, since $|f(c)| < \infty$, we have

$$\int_{\mathbb{R}} f^+ d\delta_c = f^+(c) < \infty \text{ and } \int_{\mathbb{R}} f^- d\delta_c = f^-(c) < \infty.$$

Thus, f is integrable over $\mathbb{R}$ and

$$\int_{\mathbb{R}} f \, d\delta_c = \int_{\mathbb{R}} (f^+ - f^-) \, d\delta_c = \int_{\mathbb{R}} f^+ d\delta_c - \int_{\mathbb{R}} f^- d\delta_c = f^+(c) - f^-(c) = f(c).$$

□

Example 7. Consider the integral

$$\int_{\mathbb{R}} [\cos(3x) + 2] \, d\delta_\pi.$$

Choose $t \in \mathbb{R}$ such that $t \geq 3$. Note that the set

$$\{x \in \mathbb{R} : \cos(3x) + 2 > t\} = \emptyset \in \mathcal{A}.$$

If $t < 3$, then $\{x \in \mathbb{R} : \cos(3x) + 2 > t\} \subseteq \mathbb{R}$. Hence $\{x \in \mathbb{R} : \cos(3x) + 2 > t\} \in \mathcal{A}$. It follows that $\cos(3x) + 2$ is $\mathcal{A}$-measurable. Additionally, $\cos(3\pi) + 2 = 1$. Theorem 3.3 can therefore be applied to this integral, yielding:

$$\int_{\mathbb{R}} [\cos(3x) + 2] \, d\delta_\pi = \cos(3\pi) + 2 = 1. \quad [5]$$

Example 8. Consider the integral

$$\int_{\mathbb{R}} e^{|x|+3} d\delta_2. \text{ [5]}$$

Take $t \in \mathbb{R}$ so that $t < 1$. The set $\{x \in \mathbb{R} : e^{|x|+3} < t\} = \emptyset \in \mathcal{A}$. Alternatively, set $t \geq 1$. In this case, $\{x \in \mathbb{R} : e^{|x|+3} < t\} \subseteq \mathbb{R}$. Consequently, $\{x \in \mathbb{R} : e^{|x|+3} < t\} \in \mathcal{A}$ and $e^{|x|+3}$ is $\mathcal{A}$-measurable. Recall that $e^{|2|+3} = e^5 < \infty$. Thus it is possible to use Theorem 3.3 in evaluating this integral. Doing so gives:

$$\int_{\mathbb{R}} e^{|x|+3} d\delta_2 = e^{|2|+3} = e^5 = 148.413159\ldots$$

References

[1] Dirac, Paul A.M. *The Principles of Quantum Mechanics.* 4 ed. *The International Series of Monographs on Physics.* 27, Hong Kong: Oxford University Press, 1991.

[2] Viaclovsky, Jeff. "Measure and Integration: Lecture 3." Measure and Integration (18.125). Massachusetts Institute of Technology. Cambridge. Fall 2003.

[3] Malliavin, Paul, L. Kay, H. Airault, and G. Letac. *Integration and Probability. Graduate Texts in Mathematics.* Harrisonburg: Springer-Verlag New York, Inc., 1995.

[4] Browder, Andrew. *Mathematical Analysis: An Introduction. Undergraduate Texts in Mathematics.* Harrisonburg: Springer-Verlag New York, Inc., 1996.

[5] Griffiths, David J. *Introduction to Quantum Mechanics.* 2 ed. Pearson Prentice Hall, 2004.

STUDENT ARTICLE

4 Widths in Graphs

Anand Oza[†]
Massachussets Institute of Technology '14
Cambridge, MA 02139
anandoza@mit.edu

Shravas Rao[‡]
Massachussets Institute of Technology '13
Cambridge, MA 02139
shravasa@mit.edu

Abstract

The P vs. NP question has long been an open problem in computer science, and consequently, there are many problems that have no known polynomial time solution. Recently, the idea of fixed-parameter tractable algorithms have been introduced, and have been very useful in finding faster algorithms for these problems. When restricted to inputs where some defined parameter of the input is fixed, these algorithms run in polynomial time in the input size. In this paper, we will present three parameters, all of which are widths in graphs. They include tree width, boolean width, and branch width. In addition, we will present their applications to finding FPT algorithms for NP-complete problems of finding the minimum vertex cover of a graph, the maximum independent set of a graph, and the minimum dominating set of a graph respectively.

4.1 Introduction

One of the most central problems in computer science is the question of whether or not P is equal to NP. Informally, P is the class of problems for which there exists an algorithm that can solve that problem in time proportional to a polynomial of the input size. On the other hand, NP is the class of problems for which a solution can be verified in time proportional to a polynomial of the input size. The question is then, whether all problems whose solutions can be verfied in polynomial time be solved in polynomial time

Through the study of this question, the concept of NP-completeness has arisen. A problem is considered to be NP-complete if it is in NP, and if a polynomial time solution to that problem would imply that $P = NP$, while proving that no such solution exists would imply that $P \neq NP$. Many problems, including SAT, Vertex Cover, Maximum Independent Set, and Minimum Dominating Set have been shown to be NP-complete. However, solving the P vs. NP problem has proven to be difficult, and consequently, so has finding fast solutions to NP-complete problems.

One compromise has been found in the concept of fixed-parameter tractability. Instead of measuring the speed of our algorithm based on just the input size, we can also introduce a parameter, that measures whatever we want. An algorithm is then fixed-parameter tractabile if the running time is polynomial in the input size, as long as we consider the parameter to be a constant. This means that the algorithm could be exponential, or even worse, in terms of the parameter. Interesting fixed-parameter tractable algorithms have been found for many NP-complete problems, improving on the naive exponential solutions while not yet being polynomial.

[†]Anand Oza is a sophomore at MIT studying some combination of mathematics, computer science, and physics. At this point in time, his focus is on computer science. He also enjoys racquet sports and Super Smash Bros. Melee.

[‡]Shravas Rao is a junior at MIT majoring in mathematics with computer science. His interests mainly lie in theoretical computer science. He also enjoys public television, politics, and indie music.

One specific type of parameter we will be studying in this paper is the width of a graph. There are many different types of widths, including tree-width, branch-width, boolean-width, and more, each trying to measure how complex a graph is. Because many of the known NP-complete problems are on graphs, these widths easily lend themselves to fixed-parameter tractable algorithms. Specifically, we will be looking at tree width, boolean width, branch width, and using them to find fixed-parameter tractable algorithms for the minimum vertex cover, minimum dominating set, and minimum dominating set problems respectively.

4.1.1 Dynamic Programming

All the algorithms we will present use a technique called "dynamic programming." Given a problem, we can sometimes break up the problem into subproblems, and then come up with a solution using the solutions to these subproblems. Often times these subproblems can be broken up into their own subproblems, and this continues until we have subproblems that cannot be broken up any further. The inputs that define a subproblem are usually referred to as a state. Additionally, many of these subproblems may be used more than once, in which case solving them again may be redundant. Dynamic programming takes advantage of these factors, storing the solutions of the subproblems so that they can be used later, and solving the subproblems in an order so that when solving one subproblem, we can assume that subproblems used in the solution have already been solved and their solutions are easily accessible.

For example, consider the problem of finding the nth Fibonacci number. This can be broken down into finding the $(n-1)$st and $(n-2)$nd Fibonacci numbers, as their sum is the nth Fibonacci number. These can be broken down further and further, until the subproblems to consider include finding the ith Fibonacci number for any non-negative integer i less than n, and where the state corresponding to each subproblem is i. For $i = 0$ and $i = 1$, these would need to be explicitly stated in the algorithm, as the corresponding subproblems cannot be broken down any further. We can then continue by solving the ith Fibonacci number, starting from 2 all the way up to n and then storing the values in a table. This way, when calculating the ith Fibonacci number, we just need to look up the table entries for the previous two subproblems.

Additionally, note that the running time of a dynamic programming algorithm is bounded above by the total number of states (or subproblems) multiplied by the maximum amount of time it takes to solve a subproblem, given that all of the subproblems that it breaks up into have already been solved.

4.2 Tree Width and Vertex Cover

4.2.1 Tree Width

To define tree width, we will first introduce the concept of a tree decomposition of a graph $G(V, E)$. A tree decomposition is denoted by (T, B), where $B = \{B_n\}_{n \in I}$ is a family of subsets of the set V of vertices of G indexed by n in some set I, and T is a tree whose nodes are also labeled by the set I. For clarity, we refer to the vertices of G as "vertices" and the vertices of T as "nodes." Additionally, we require that a tree decomposition of G be so that: every vertex of G is contained in B_i for some $i \in I$, for every edge e in G, the two adjacent vertices are contained in a set B_i for some i, and for every vertex v, the subgraph of T induced by the set of nodes i in T in which v in conatined in B_i must be connected.

The tree width of a particular tree decomposition, (T, B), is the size of the largest set B_i, minus 1. The *tree width* of a graph G is the minimum tree width of any tree decomposition of G; any tree decomposition of minimal tree width is called an *optimal tree decomposition*

Informally, the tree width of a graph G represents how close the graph is to being a tree. For example, the tree-width of a graph that is a tree is just 1, as we can place each the two vertices adjacent to an edge of the graph in a different node of the tree decomposition. The edges of the tree decomposition exist between two nodes of the tree decomposition if they both contain a vertex v, and one of the nodes contains v's parent. However, the tree width of a complete graph on n vertices is $n - 1$, as all the vertices need to be in the same set B_i for some i. Otherwise, by the final condition on a tree decomposition, there would exist a pair of vertices not in the same node

of the tree decomposition, violating the second condition. Since many problems can be solved more easily on trees rather than on graphs in general, we can generalize the ideas used for these algorithms to algorithms on a tree decomposition of a graph.

4.2.2 Vertex Cover

A *vertex cover* C of a graph $G(V, E)$ is a subset of the vertices such that every edge of the graph is adjacent to a vertex in C. The minimum vertex cover problem is that of finding a vertex cover C of a given graph G of minimum size.

We will describe an algorithm whose running time is $\mathcal{O}(2^{k^4}n)$, given in [4], where k is the tree width of the input graph, and n is the number of vertices. This is therefore fixed-parameter tractable in tree width.

4.2.2.1 Algorithm

Our algorithm works in two main steps. First, we find the optimal tree decomposition of the input graph G. There exist an algorithm to find the optimal tree decomposition in time $O(2^{k^4}n)$ by [1], where k is the tree-width of the graph G. For our purposes, we will assume an optimal tree decomposition, (T, B). Additionally, the number of nodes in this optimal tree decomposition is $O(n)$, where n is the number of vertices in G.

Once we are given a tree decomposition, we can find a minimum vertex cover by preforming a dynamic programming algorithm. We first start by rooting the tree at some vertex, r. Then for each node n in the graph T, we can define the set of vertices V_n of G, which is the set of vertices contained in the set B_i for any node i in T with n as an ancestor. We can also define the induced subgraph G_n of G by including only the vertices V_n. Our subproblem is then given a node n from T and a subset S of B_n, to find the size of the smallest vertex cover of G_n, that contains S, but does not cotain any other vertex in B_n. We will store this value in $f(n, S)$, using ∞ if no such vertex cover exists. Note that the minimum value stored in $f(r, B'_r)$, for any B'_r that is a subset of B_r, is the size of the minimum vertex cover, as G_r is the same as G.

We can solve the subproblems $f(n, S)$ in postorder of the vertices in T. This allows us to assume that when we attempt to calculate the value of $f(n, S)$, the values stored in $f(n', S')$, where n' is a child of n and for any possible subset of $B_{n'}$, S', have already been calculated. Then we can use the values of $f(n', S')$ to help us solve the subproblem $f(n, S)$.

Now, we can describe how to solve the subproblem $f(n, S)$. If n is a leaf of T, then V_n is equal to B_n. Therefore, $f(n, S)$ has a valid corresponding vertex cover, iff S is a vertex cover of G_n, in which case, the size is $|S|$. Otherwise, no valid vertex cover exists and we store ∞.

Now, consider the case where n is not a leaf. First, we have to make sure that for every edge between two vertices in B_n, one endpoint is contained in S. Otherwise, there is no vertex cover that contains S, but not the other vertices in B_n, and we can store ∞ in $f(n, S)$. If this is true, then we visit each child, one by one.

For each child, n', we iterate over all subsets S' of B'_n. However, in the case there are vertices contained in both B_n and B'_n, we will only consider those sets for which $S' \cap B_n = S \cap B_{n'}$. Essentially, we want the assignments of S' and S to agree with other. We then pick the set S' so that $f(n', S') - |S' \cap B_n|$ is minimized, and add this value to a running sum. We subtract $|S' \cap B_n|$, as these are already counted with the subset $|S|$. Finally, we add $|S|$ to the running sum to get the solution for the subproblem, and store this value in $f(n, S)$. Note that if for any child n', and for all valid sets S', the value stored in $f(n, S')$ is ∞, then we can not assign a value to $f(n, S)$ as no corresponding vertex cover exists.

Finally, we iterate over all $f(r, B'_r)$ for any B'_r that is a subset of B_r, and return the minimum of these as our result.

4.2.2.2 Proof of Correctness

To show that our algorithm is correct, we must show that for each pair (n, S), the value stored in $f(n, S)$ is correct. In other words, we need to show that there is a vertex cover of G_n that includes S, but not the other vertices in B_n, of size $f(n, S)$, and that no smaller such vertex cover exists. This would then imply that our final answer is also correct.

We can start by showing such a vertex cover exists, using induction. The base case is where n is a leaf. In this case, the vertices of G_n are B_n, and the algorithm requires that any edge between vertices in B_n be covered by S. Therefore, S is a vertex cover of G_n.

Otherwise, we can assume that such vertex covers exist for the children of n. For each child, n_i, let S_i be a subset of $B_{n'}$ so that the condition $S' \cap B_n = S \cap B_{n'}$ holds, and that $f(n, S') - |S' - S|$ is minimized. Then the union of vertex covers of G_{n_i} with size $f(n_i, S_i)$, along with S, create a vertex cover of G_n of minimum size that contains S, but not the other vertices in B_n.

Now we prove that no smaller vertex cover can exist by induction. Again, our base case is the case where n is a leaf. If S is a valid vertex cover of G_n, then no vertex cover containing S can be smaller than $f(n, S)$. Otherwise, there are no vertex cover not containing the other vertices of B_n.

If n is not a leaf, then consider a vertex cover C of G_n containing S and not any other vertex in B_n. Then for every child n' of n, we also have a vertex cover $C_{n'}$ of $G_{n'}$ using the vertices from C also in $G_{n'}$. However, by our induction hypothesis, each $C_{n'}$ can be no smaller than $f(n', B_{n'} \cap C_{n'})$. However, because C is equal to the union of $C_{n'}$ for all children n' with S, and S is equal to $C \cap B_n$, the size of C cannot be smaller than $f(n, S)$. This completes the proof.

4.2.2.3 Runtime Analysis

Our algorithm considers $\mathcal{O}(n2^k)$ states. For each state, we consider all the states corresponding to each child to solve this subproblem. However, because this is a tree, the total number of parent-child pairs is $\mathcal{O}(n)$, this results in a running time of $\mathcal{O}(n2^{2k})$. Because calculating the optimal tree decomposition takes $\mathcal{O}(n2^{k^4})$ time, the overall running time is $\mathcal{O}(n2^{k^4})$, which is fixed-parameter tractable in tree-width.

4.3 Boolean Width and Independent Set

4.3.1 Boolean Width

To define boolean width, we will introduce a decomposition tree of a graph, $G(V, E)$, different from that used to define tree width. We denote this decomposition tree as (T, δ), where T is an arbitrary tree with $|V|$ leaves, and δ is a bijective mapping from the leaves of T to the set V. Note that removing an edge e from T separates T into two components, and therefore partitions the vertices of G (which correspond to the leaves of T) into two disjoint subsets, V_1 and V_2. The idea behind this decomposition tree is to start with the individual vertices of G in the leaves, and slowly group them together as we move up the tree to the root. This gives an organized manner in which to consider only certain subsets of V.

Now, for each separation, we can define a boolean dimension. Then, the boolean width of a decomposition tree is the maximum boolean dimension over its separations, and the *boolean width* of a graph is the minimum boolean width over all decomposition trees. Note that there are more seperations in a graph than defined by a decomposition tree.

For a given separation, let A be V_1, and $\overline{A}$ be the rest of the graph. Then, let $N(X)$ be the set of all vertices that share an edge with at least one vertex of X. Then let $U(A)$ be the set of all sets $N(X) \cap \overline{A}$, where X is a subset of A. Note that two different subsets of A, X and X', can be so that $N(X) \cap \overline{A} = N(X') \cap \overline{A}$. The boolean dimension of the separation is $\log_2 |U(A)|$.

An optimal tree decomposition in terms of boolean width can be helpful if in a cut $\{A, \overline{A}\}$, we can solve the problem in the subgraph defined by the vertices in A, and then finish the problem based on only the set of neighbors in $\overline{A}$ of some set in A. For example, if X and X' are sets of vertices contained in A, but both have the same set of neighbors in $\overline{A}$, then a vertex in $\overline{A}$ does not share an edge with a vertex in X if and only if it does not share an edge with a vertex in X'. For some problems, this may allow us to consider how we want to treat vertices from $\overline{A}$, independent of our choice of X and X'.

4.3.2 Maximum Independent Set

An *independent set* I of a graph $G(V, E)$ is a subset of the vertices such that no two vertices are adjacent. The minimum independent set problem is that of finding an independent set I of a given graph G of minimum size.

We will describe a fixed-parameter tractable algorithm, given in [2], whose running time is $\mathcal{O}(nk^2 2^{2k})$ where n is the number of vertices of G and k is the boolean width. This is therefore, fixed-parameter tractable in boolean width.

4.3.2.1 Algorithm

Again the algorithm includes two major steps. First, we need to have an optimal (minimum boolean width) decomposition tree, T. By [2], computing an optimal decomposition tree is fixed parameter tractable in terms of boolean width. Let the boolean width be k. Note that this tree is rooted.

Additionally, for each cut $\{A, \overline{A}\}$ described by the decomposition tree, we define a list $L(A)$ of representative subsets of A such that no two elements of $L(A)$ have the same neighborhood in $\overline{A}$, but every subset of A has the same neighborhood as some element of $L(A)$. If we arrange subsets of A into equivilince classes so that subsets in each equivilence class have the same neighborhood in A, then $L(A)$ contains a representative from each equivilence class. Note that the size of $L(A)$ is at most 2^k.

In [2], a data structure is presented that creates a list $L(A)$ in time $\mathcal{O}(nk^2 2^{2k})$, and allows us to find a representative of a subset of A in time $\mathcal{O}(k)$. For our purposes, we will be assuming such a data structure.

The states for our dynamic programming algorithm are of the form (n, S), where n is a n of T and S is an element of $L(A)$, where $\{A, \overline{A}\}$ is the cut associated with the parent edge of n (unless n is the root, in which case S can only be V). We will store in $f(n, S)$ the maximum size of an independent set $I \subseteq A$ such that I shares the same neighborhood as S in $\overline{A}$ (S is the representative of I). Our final answer to the maximum independent set problem is the value in $f(root, S)$.

The algorithm visits the vertices of the decomposition tree in post-order. This way, when we evaluate $f(n, S)$ for some vertex n, we can assume we know the value of $f(n', S')$ where n' is a child of n. If n is a leaf, then this value stored is just the size of S, as S contains either 0 or 1 vertices.

Now, consider finding the value of $f(n, S)$ when n is not a leaf. Let n_1 and n_2 be the children of n, and then let A_1 and A_1 be the cuts that correspond to these vertices respectively. Then iterate over all pairs S_1 and S_2 from the set $L(A_1)$ and $L(A_2)$ respectively. Then we check that there are no edges between S_1 and S_2, and that S is the representative of $S_1 \cup S_2$ in $L(A)$. If this is true, then there is an independent set that shares the same neighborhood as S in $\overline{A}$ of size $f(n_1, S_1) + f(n_2, S_2)$. We choose the maximum of these to store in $f(v, S)$.

Finally, we can return $f(root, S)$ as our solution.

When we are done, the answer is in $f(root, S)$ (the last entry to be updated), as stated earlier.

4.3.2.2 Proof of Correctness

In order to show $f(n, S)$ is indeed the size of the maximum independent set $I \subseteq A$ whose representative is S, as desired, we must show that such a set of size $f(n, S)$ exists, and that $f(n, S)$ is an upper bound. We will prove this by showing that the value stored in $f(n, S)$ is correct for all entries, using induction on the vertices of the tree in postorder.

For the base case, we consider the leaves of the tree. For a leaf n, we have $f(n, S) = |S|$ and there is only one vertex in A, so the maximum independent set with the same neighborhood as S is just S.

Now, consider $f(n, S)$ where n is not a leaf. We must first show that for each internal node v and associated representative S, there is an independent set $I \subseteq A$ whose representative is S, of size $f(v, S)$. For a given state (v, S), we know there exist (v_1, S_1) and (v_2, S_2) such that $f(v, S) = f(v_1, S_1) + f(v_2, S_2)$, and these two states were the ones used to update $f(v, S)$, according to the algorithm. By our inductive hypothesis, there are independent sets I_1 and I_2 in A_1 and A_2, respectively, of sizes $f(v_1, S_1)$ and $f(v_2, S_2)$, respectively. We claim $I = I_1 \cup I_2$ is an independent set, and its representative is S. Suppose, for contradiction, that there is an edge between I_1 and I_2. Because I_1 and S_1 have the same neighborhood in $\overline{A}_1$, which includes A_2, there must be an edge between S_1 and I_2. Similarly, because I_2 and S_2 have the same neighborhood in $\overline{A}_2$, which includes A_1, there must be an edge between S_1 and S_2. This is a contradiction, because our algorithm checks that there are no edges between S_1 and S_2. Therefore, I is an independent set.

Furthermore, because $\overline{A}_1$ contains $\overline{A}$, we know I_1 and S_1 have the same neighborhood in $\overline{A}$, and the same is true for I_2 and S_2. Therefore, $I = I_1 \cup I_2$ has the same neighborhood as $S = S_1 \cup S_2$, so I's representative is S.

Next, we must show that for each internal node v and associated representative S, $f(n, S)$ is an upper bound on the size of any independent set in A whose representative is S. Consider an independent set I with representative S. Let n_1 and n_2 be the two children of n, and define $I_1 = I \cap A_1$ and $I_2 = I \cap A_2$. Because I_1 and I_2 are subsets of I, they are also independent sets, and because they are subsets of A_1 and A_2, respectively, we can let S_1 and S_2 be their representatives. There are no edges between S_1 and S_2, because I_1 and S_1 have the same neighborhood in $\overline{A}_1$ and I_2 and S_2 have the same neighborhood in $\overline{A}_2$. Furthermore, S is the representative of $S_1 \cup S_2$, because $S_1 \cup S_2$ has the same neighborhood as I in $\overline{A}$ (because S_1 has the same neighborhood as I_1 in $\overline{A}_1 \supseteq \overline{A}$, and same for 2). Therefore, by our inductive hypothesis and the algorithm, we know $|I_1| \leq f(n_1, S_1)$ and $|I_2| \leq f(n_2, S_2)$. Because $I = I_1 \cup I_2$, we know $|I| \leq |I_1| + |I_2| \leq f(n_1, S_1) + f(n_2 + S_2) \leq f(n, S)$, as desired.

4.3.2.3 Runtime Analysis

Our algorithm considers $\mathcal{O}(n2^k)$ states - there are $\mathcal{O}(n)$ vertices, and for each vertex there are at most $\mathcal{O}(2^k)$ representatives to consider. For each state, we look at $\mathcal{O}(2^{2k})$ pairs of additional states, as we have already decided on the vertices of the states we want to consider. Checking that there are no edges between the selected representatives, S_1 and S_2 takes $\mathcal{O}(n^2)$ time, and checking that S is the representative of $S_1 \cup S_2$ takes $\mathcal{O}(k)$. Therefore, the overall running time is $\mathcal{O}(n^2 k 2^{3k})$. Through a tighter analysis found in [2], the running time of this algorithm can be shown to be $\mathcal{O}(nk^2 2^{2k})$.

4.4 Branch Width and Dominating Set

4.4.1 Branch Width

As with boolean width, the definition of branch width is based on decomposition trees. However, in the definition of branch width, we map the leafs of the decomposition tree to edges, rather than to vertices.

Specifically, a branch decomposition of a graph $G(V, E)$, is denoted by (T, τ), where T is an arbitrary tree with $|E|$ leaves and τ is a bijective mapping from the leaves of T to the set E. As with boolean width, we require that every node of T have degree either 1 or 3. Note that removing an edge e from T separates T into two components, and therefore partitions the edges of G (which correspond to the leaves of T) into two disjoint subsets, E_1 and E_2, whose union is E. If $U(E_1)$ is the set of vertices in common between the two edge sets E_1 and E_2, then the "width" of this seperation is $|U(E_1)|$. The branch width of a decomposition (T, τ) is the maximum "width" over all separations created by removing a single edge of T.

The *branch width* of a graph is the minimum width over all the branch decompositions of the graph. A decomposition tree is the useful, because it allows us to consider only certain vertices at a time. If e, and e', and e'' are edges of T so that e is a child of e', and e' is a child of e'', then if a vertex v is in both $U(E_1)$ and $U(E_1'')$, then a vertex must also be in $U(E_1')$. In general, as we move up a tree, once a vertex no longer appears in $U(E_1)$, where E_1 is the edge set we are considering, it will never appear again. This allows us to consider each $U(E_1)$ in comparison which only a set in a child, or in a parent.

4.4.2 Minimum Dominating Set

A *dominating set* D of a graph $G(V, E)$ is a subset of the vertices such that every vertex of the graph is either in D or adjacent to a vertex in D. The minimum dominating set problem is that of finding a dominating set D of a given graph G of minimum size.

We will describe a fixed parameter tractable algorithm, given in [3], running time is $\mathcal{O}(3^{1.5 \cdot \ell} m)$, where ℓ is the branch width of the input graph, and m is the number of edges. This is therefore, fixed-parameter tractable in branch width.

4.4.2.1 Algorithm

As with the previous two algorithms, this one works in two main steps: first, we compute an optimal branch decomposition of the reduced graph. There is a polynomial time algorithm for this, by [3].

Once we are given a branch decomposition, we solve the dominating set problem on the graph using dynamic programming and the branch decomposition. We will describe this part of the algorithm, in more detail.

Let (T', τ) be a branch decomposition of G. The first step is to turn T' into a rooted tree, T, by adding edges and vertices, so that for every vertex in T', there is a parent of that vertex in T. To do so, we pick an arbitrary edge $\{x, y\}$ of T' and insert a new vertex v in the middle of the edge. Then, we connect v to a new vertex r. Let $T = T' \cup \{v, r\}$, and let r be the root of T.

We then define an order function, $\omega'(e)$, which takes an edge e of the tree T' and returns a subset of the vertices of G. If removing the edge e from T' splits the leafs of T', and therefore the edges of G, into the two edge sets E_1 and E_2, then $\omega'(e)$ is the set of vertices the two edge sets have in common. For T, we keep the same values of ω for edges also in T', but for the three new edges, we assign $\omega(\{x, v\}) = \omega(\{y, v\}) = \omega'(\{x, y\})$, and $\omega(\{v, r\}) = \emptyset$. Note that $|\omega(e)|$ in a given decomposition tree is always less than or equal to the branch width of the decomposition tree, by the definition of branch width.

Finally, we define G_e to be the subgraph of G induced by the edges of G corresponding to the leaves of T whose path to the root, r, includes e (i.e., the leaves that are descendants of e).

Now we will introduce the idea of a coloring, which will later be used to define a state for the dynamic programming algorithm. For every edge e of T, we can color the vertices of $\omega(e)$ with 1, 0, or $\hat{0}$. A color of 1 indicates that the vertex is in the dominating set, a color of 0 indicates the vertex is not in the dominating set, and has already been dominated by a vertex currently in the dominating set, and a color of $\hat{0}$ indicates we have not yet decided whether it should be 1 or 0.

For our dynamic programming algorithm, we have as our state an edge e of T along with a coloring c of the vertices of $\omega(e)$, colored using using 1, 0, and $\hat{0}$. For each state, we store in $f(e, c)$ the minimum cardinality of a set $D_e \subseteq V(G_e)$ that is a dominating set of G_e, so that this set agrees with the coloring c. If no such set exists, we store ∞. D_e must be such that all vertices in $\omega(e)$ colored by 1 are contained in D_e, all vertices by 0 in $\omega(e)$ are dominated by D_e (and are not in D_e), and all vertices of G_e not in $\omega(e)$ are dominated by D_e. Note that the value stored in $f(\{v, r\}, \emptyset)$ is the desired result, as $G_{\{v,r\}}$ is just the graph G, and $\omega(\{v, r\})$ is empty.

For a nonleaf edge e of T, it has two child edges e_1, e_2. We say a coloring c of $\omega(e)$ can be *formed* from colorings c_1 and c_2, of $\omega(e_1)$ and $\omega(e_2)$, respectively, if the following four conditions hold:

1. For every $u \in \omega(e) - \omega(e_2)$, $c(u) = c_1(u)$.
2. For every $u \in \omega(e) - \omega(e_1)$, $c(u) = c_2(u)$.
3. For every $u \in \omega(e) \cap \omega(e_1) \cap \omega(e_2)$:
   (a) $c(u) = 0$ only if neither of $c_1(u)$ and $c_2(u)$ are 1.
   (b) $c(u) = \hat{0}$ only if $c_1(u) = c_2(u) = \hat{0}$.
   (c) $c(u) = 1$ only if $c_1(u) = c_2(u) = 1$.
4. For every $u \in (\omega(e_1) \cup \omega(e_2)) - \omega(e)$, one of the following holds:
   (a) $c_1(u) = c_2(u) = 0$.
   (b) or $c_1(u) = c_2(u) = 1$.
   (c) or $c_1(u) = 0$ and $c_2(u) = \hat{0}$.
   (d) or $c_1(u) = \hat{0}$ and $c_2(u) = 0$.

These conditions come about from the fact that if a vertex is colored 1 or 0 in at least two of $\omega(e)$, $\omega(e_1)$, and $\omega(e_2)$, then its color must stay constant over all sets it belongs to. If it is colored $\hat{0}$, then a bit more leeway is given, as its inclusion in the dominating set is still uncertain.

Now we will present the algorithm. Consider finding the value of $f(e, c)$. If e is adjacent to a leaf in T, then we just check if c is valid. Specifically, we just check to see if every vertex colored 0 is adjacent to some vertex colored 1. If so, we store the number of vertices of c are colored 1. Otherwise, we store ∞.

Otherwise, let e_1 and e_2 be the child edges of e. We iterate over all possible colorings, c_1 and c_2 of $\omega(e_1)$ and $\omega(e_2)$ respectively, in which c can be formed from c_1 and c_2 and so that neither $f(e_1, c_1)$ nor $f(e_2, c_2)$ are ∞. If no such pair exists, then we store ∞ in $f(e, c)$.

Given colorings c_1 and c_2 of $\omega(e_1)$ and $\omega(e_2)$ in which c can formed from c_1 and c_2, we can create a dominating set of G_e from the dominating sets of G_{e_1} and G_{e_2} corresponding to $f(c, e_1)$ and $f(c, e_2)$ respectively, along with what the coloring c indicates. The size of this dominating set is $f(e_1, c_1) + f(e_2, c_2) - \#_1(\cap\omega(e_1) \cap \omega(e_2), c_1)$, where $\#_1(X, c)$ the number of vertices in set X marked with "1" in the coloring c (the last term keeps us from overcounting vertices). The minimum such value is what we store in $f(e, c)$.

Finally, we can return $f(\{v, r\}, \emptyset)$ as our result.

4.4.2.2 Proof of Correctness

We can prove correctness of the value stored in $f(e, c)$ by induction on the vertices, in postorder. Because we are visiting edges in postorder, we can assume that when e is not adjacent to a leaf, any $f(e_1, c_1)$ and $f(e_2, c_2)$ is assigned the correct value, where e_1 and e_2 are child edges of e. If e is adjacent to a leaf, there is only one step to consider, which is obviously true.

If there exist children e_1, e_2 of an edge e, then the vertex set $\omega(e_1)$ cuts off the graph G_{e_1} from the rest of G, so we only have to worry about the vertices in $\omega(e_1)$. The same applies to $\omega(e_2)$ and G_{e_2}. Therefore, the union of a domiating set of G_{e_1} and the dominating set of G_{e_2} will create a dominating set of G_e, as long as the shared vertices between G_{e_1} and G_{e_2} agree. Therefore, as long as our coloring c, c_1 and c_2 satisfy the conditions listed and neither $f(e_1, c_1)$ nor $f(e_2, c_2)$ are ∞, there exist dominating sets of size $f(e_1, c_1)$ and $f(e_2, c_2)$ in G_{e_1} and G_{e_2} respectively that can be combined into a dominating set of G_e. Note that the expression used to assign $f(e, c)$ come from inclusion-exclusion, since some vertices are counted twice in the term $f(e_1, c_1) + f(e_2, c_2)$. Additionally, because of the conditions in the definition of form, the last term would be the same if c_1 were replaced by c_2.

4.4.2.3 Runtime Analysis

To analyze the runtime of this algorithm, we first consider runtime in calculating the value of each $f(e, c)$. For each state, there are at most $\mathcal{O}(3^{2\ell})$ possible pairs of c_1 and c_2 to consider. Additionally, there are $\mathcal{O}(3^{\ell}m)$ states to consider, for an overall running time of $\mathcal{O}(3^{3\ell}m)$. Therefore, minimum dominating set is fixed-paramter tractable in branch width. A tigheter analysis given in [3] results in a faster runtime of $\mathcal{O}(3^{1.5\ell}m)$.

4.5 Conclusion

We have described three widths of graphs tree width, boolean width, and branch width, along with their applications to three NP-complete problems minimum vertex cover, maximum independent set, and minimum dominating set respectively. In particular, we presented an algorithm for each problem using the appropriate width as a parameter, showing that the problems are fixed parameter tractable with respect to the appropriate width.

However, there are many more widths of graphs, and many more aspects of these widths to explore. For instance, there has been much work on bounding the the widths of graphs, either in terms of the number of vertices of the graph, or even in terms of other widths. In some cases, certain classes of graphs, such as planar graphs, have even tighter bounds on the width of the graph. Additionally, the problem of calculating the width of a graph along with its decomposition is an interesting problem by itself. Finally, many of these width have only very recently been introduced, and their full potential may not yet be realized.

References

[1] Hans L. Bodlaender. A linear-time algorithm for finding tree-decompositions of small treewidth. *SIAM J. Comput.*, 25:1305–1317, December 1996.

[2] Binh-Minh Bui-Xuan, Jan Arne Telle, and Martin Vatshelle. Boolean-width of graphs. *Theoretical Computer Science*, 412(39):5187 – 5204, 2011.

[3] Fedor V. Fomin and Dimitrios M. Thilikos. Dominating sets in planar graphs: branch-width and exponential speed-up. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '03, pages 168–177, Philadelphia, PA, USA, 2003. Society for Industrial and Applied Mathematics.

[4] David Karger. 6.854 problem set 8 solutions. Problem 4(b), November 2011.

FACULTY FEATURE ARTICLE

5
Tea Time in Princeton

Professor Paul Bourgade[†]
Harvard University
Cambridge, MA 02138
bourgade@math.harvard.edu
http://www.math.harvard.edu/~bourgade.html

He said: "That's the form factor for the pair correlation of eigenvalues of random Hermitian matrices!"

This note is about who "He" is, what "That" is, and why you should never miss tea time.

Since the seminal work of Riemann, it is well-known that the distribution of prime numbers is closely related to the behavior of the ζ function. Most importantly, it was conjectured in [13] that all its (non-trivial) zeros are aligned[1], and Hilbert and Pólya put forward the idea of a spectral origin for this phenomenon.

I spent two years in Göttingen ending around the begin of 1914. I tried to learn analytic number theory from Landau. He asked me one day : "You know some physics. Do you know a physical reason that the Riemann hypothesis should be true?" This would be the case, I answered, if the nontrivial zeros of the ξ-function were so connected with the physical problem that the Riemann hypothesis would be equivalent to the fact that all the eigenvalues of the physical problem are real.

George Pólya, correspondence with Andrew Odlyzko, 1982.

Despite the lack of progress concerning the horizontal distribution of the zeros (i.e. all their real parts being supposedly equal), some support for the Hilbert-Pólya idea came from the vertical distribution, i.e. the distribution of the gaps between the imaginary parts of the non-trivial zeros. Indeed, in 1972, the number theorist Hugh Montgomery evaluated the pair correlation of these zeros, and the mathematical physicist Freeman Dyson realized that they exhibit the same repulsion as the eigenvalues of typical large random Hermitian matrices. In this expository note, we aim at explaining Montgomery's result, placing emphasis on the common points with random matrices. These statistical connections have since been extended to many other L-functions (e.g. over function fields, cf. [12]); for the sake of brevity we only consider the Riemann zeta function, and refer for example to [8] for many other connections between analytic number theory and random matrices.

5.1 Independent random points

As a first step towards the repulsion between some particles, eigenvalues or zeros of the zeta function, we wish to understand what happens when there is *no* repulsion, in particular for *independent*

[†]Paul Bourgade was born in Dax, he earned his BSc from École Polytechnique and received his PhD from Université Pierre et Marie Curie, in Paris. He is a Benjamin Peirce fellow at Harvard University, and his mathematical interests include probability and analytic number theory.

[1]For a definition of the Riemann zeta function and the Riemann hypothesis, see the beginning of Section 2.

random points. For this, consider the following natural question.

Choose n independent and uniform points on the interval $[0,1]$. What is the typical spacing between two successive such points?

A good way to make this question more precise is to assume that amongst these points $x_1, \dots, x_n$, we label one, say x_1, and we consider the probability that it has no right-neighbor up to distance δ. Denoting $\chi(I)$ the number of x_i's in an interval I, the probability of such an event is

$$\int_0^1 \mathbb{P}(\chi((y, y+\delta]) = 0 \mid x_1 = y)\mathrm{d}y,$$

because x_1 is uniformly distributed. Now, as all the x_i's are independent, the integrand is also (when $y + \delta < 1$)

$$\mathbb{P}(\cap_{i=2}^n \{x_i \notin (y, y+\delta]\}) = \prod_{i=2}^n \mathbb{P}(x_i \notin (y, y+\delta]) = (1-\delta)^{n-1}.$$

Choosing $\delta = \frac{u}{n}$ and considering the limit $n \to \infty$, we get that the probability that the gap

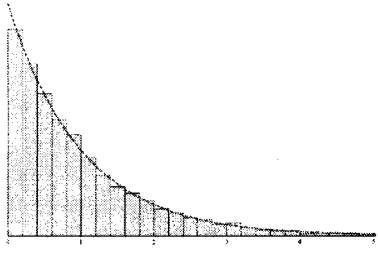


Figure 5.1: Histogram of 10^5 nearest-neighbor spacings (i.e. $n\Delta$). Dashed: the rescaled e^{-u} curve.

between x_1 and its right neighbor is greater than $\frac{u}{n}$ converges to e^{-u}. More generally, denoting by Δ the gap between x_1 and its right-neighbor, we obtain that, for any $0 < a < b$,

$$\mathbb{P}(n\Delta \in [a,b]) \underset{n\to\infty}{\longrightarrow} \int_a^b e^{-u}\mathrm{d}u\,. \tag{5.1}$$

Another way to quantify the microscopic structure of these independent points consists in looking at the following statistics, $r(f,n) = \frac{1}{n}\sum_{1\le j,k\le n, j\neq k} f(n(x_j - x_k))$, for a generic test function f. The reader will easily prove the following asymptotics:

$$\mathbb{E}\left(r(f,n)\right) \underset{n\to\infty}{\longrightarrow} \int_{\mathbb{R}} f(y)\mathrm{d}u. \tag{5.2}$$

This limiting exponential distribution (5.1) and the pair correlation (5.2) appear universally, i.e. when the sampled points are sufficiently close to independence, no matter which distribution they have [2]. It is a natural question whether this remains valid for other random points, and we will explain what happens when considering the ζ zeros with large imaginary part or the eigenvalues of random matrices. The gap statistics will be very different, both for the former (Section 2) and the latter (Section 3), for which a common type of correlations appears in the limit. The following sections are widely independent.

[2] For example, the reader could consider independent points with strictly positive density with respect to the uniform measure on $[0,1]$, and he would obtain an exponential law in the limit as well.

5.2 The pair correlation of the ζ zeros.

In this section, we state some elementary properties of the Riemann zeta function, mentioning along the way a formal analogy between the ζ zeros and the eigenvalues of the Laplacian on some symmetric spaces. We then come to more quantitative estimates through Montgomery's result on the repulsion between the ζ zeros.

For $\sigma = \Re(s) > 1$, the Riemann zeta function can be defined as a Dirichlet series or an Euler product:

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p \in \mathcal{P}} \frac{1}{1 - \frac{1}{p^s}},$$

where $\mathcal{P}$ is the set of all prime numbers. The second equality is a consequence of the expansion $(1 - p^{-s})^{-1} = \sum_{k \geq 0} p^{-ks}$ and uniqueness of factorization of integers into prime numbers. Remarkably, as proved in Riemann's original paper, ζ can be meromorphically extended to $\mathbb{C} - \{1\}$, and this extension satisfies a functional equation (see e.g. [15] for a proof): writing $\xi(s) = \pi^{-s/2}\Gamma(s/2)\zeta(s)$, we have

$$\xi(s) = \xi(1 - s).$$

Consequently, the zeta function admits trivial zeros at $s = -2, -4, -6, \ldots$ corresponding to the poles of $\Gamma(s/2)$. All the other zeros are confined in the critical strip $0 \leq \sigma \leq 1$, and they are symmetrically positioned about the real axis and the critical line $\sigma = 1/2$. The Riemann Hypothesis states that all of this *non-trivial* zeros are exactly on the line $\sigma = 1/2$.

Trace formulas. The first similarity between the zeta zeros and spectral properties of operators occurs when looking at linear statistics. Namely, we state the Weil explicit formula concerning the ζ zeros and Selberg's trace formula for the Laplacian on surfaces with constant negative curvature.

First consider the Riemann zeta function. For a function f: $(0, \infty) \to \mathbb{C}$, define its Mellin transform $F(s) = \int_0^\infty f(s) x^{s-1} \mathrm{d}x$. Then the inversion formula (where σ is chosen in the fundamental strip, i.e. where the image function F converges)

$$f(x) = \frac{1}{2\pi \mathrm{i}} \int_{\sigma - \mathrm{i}\infty}^{\sigma + \mathrm{i}\infty} F(s) x^{-s} \mathrm{d}s$$

holds under suitable smoothness assumptions, in a similar way as the inverse Fourier transform. Hence, for example,

$$\sum_{n=2}^{\infty} \Lambda(n) f(n) = \sum_{n=2}^{\infty} \Lambda(n) \frac{1}{2\pi \mathrm{i}} \int_{2 - \mathrm{i}\infty}^{2 + \mathrm{i}\infty} F(s) n^{-s} \mathrm{d}s = \frac{1}{2\pi \mathrm{i}} \int_{2 - \mathrm{i}\infty}^{2 + \mathrm{i}\infty} \left(-\frac{\zeta'}{\zeta} \right)(s) F(s) \mathrm{d}s,$$

where Λ is Van Mangoldt's function[3]. To derive the above formula, we use that $-\frac{\zeta'}{\zeta}(s) = \sum_{n \geq 2} \frac{\Lambda(n)}{n^s}$, which is obtained by deriving the formula $-\log \zeta(s) = \sum_{\mathcal{P}} \log(1 - p^{-s})$. Now, changing the line of integration from $\Re(s) = 2$ to $\Re(s) = -\infty$, all trivial and non-trivial poles (as well as $s = 1$) are crossed, leading to the following formula,

$$\sum_{\rho} F(\rho) + \sum_{n \geq 0} F(-2n) = F(1) + \sum_{p \in \mathcal{P}, m \in \mathbb{N}} (\log p) f(p^m),$$

where the first sum is over non-trivial zeros counted with multiplicities. When replacing the Mellin transform by the Fourier transform, the above formula linking linear statistics of zeros and primes takes the following form, known as the Weil explicit formula.

[3] $\Lambda(n) = \log p$ if $n = p^k$ for some prime p, 0 otherwise.

Theorem. Let h be even, analytic on $|\Im(z)| < 1/2 + \delta$, bounded, and decreasing as $h(z) = O(|z|^{-2-\delta})$ for some $\delta > 0$. Here, the sum is over all γ_n's such that $1/2 + i\gamma_n$ is a non-trivial zero, and $\hat{h}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} h(y)e^{-ixy}dy$:

$$\sum_{\gamma_n} h(\gamma_n) - 2h\left(\frac{i}{2}\right) = \frac{1}{2\pi}\int_{\mathbb{R}} h(r)\left(\frac{\Gamma'}{\Gamma}\left(\frac{1}{4}+\frac{i}{2}r\right) - \log\pi\right)dr - 2\sum_{p\in\mathcal{P}m\in\mathbb{N}} \frac{\log p}{p^{m/2}}\hat{h}(m\log p). \quad (5.3)$$

In a very distinct context holds a similar relation, the Selberg's trace formula. In one of its simplest manifestations, it can be stated as follows. Let $\Gamma\backslash\mathbb{H}$ be a quotient of the Poincaré half-plane, where Γ is a subgroup of $PSL_2(\mathbb{R})$, the orientation-preserving isometries of $\mathbb{H} = \{x + iy, y > 0\}$ endowed with the metric

$$(ds)^2 = \frac{(dx)^2 + (dy)^2}{y^2}. \quad (5.4)$$

The Laplace-Beltrami operator $\Delta = -y^2(\partial_{xx} + \partial_{yy})$ is self-adjoint with respect to the invariant measure associated to (5.4), $d\mu = \frac{dxdy}{y^2}$, i.e. $\int v(\Delta u)d\mu = \int(\Delta v)ud\mu$, so all eigenvalues of Δ are real and positive. If $\Gamma\backslash\mathbb{H}$ is compact, the spectrum of Δ restricted to a fundamental domain $\mathcal{D}$ of representatives of the conjugation classes is discrete, noted $0 \leq \lambda_0 < \lambda_1 < \dots$ To state Selberg's trace formula, we need, as previously, a function h analytic on $|\Im(z)| < 1/2 + \delta$, even, bounded, and decreasing as $h(z) = O(|z|^{-2-\delta})$, for some $\delta > 0$.

Theorem. Under the above hypotheses, setting $\lambda_k = s_k(1 - s_k)$, $s_k = 1/2 + ir_k$, then

$$\sum_{k=0}^{\infty} h(r_k) = \frac{\mu(\mathcal{D})}{2\pi}\int_{-\infty}^{\infty} rh(r)\tanh(\pi r)dr + \sum_{p\in\mathcal{P},m\in\mathbb{N}^*} \frac{\ell(p)}{2\sinh\left(\frac{m\ell(p)}{2}\right)}\hat{h}(m\ell(p)), \quad (5.5)$$

where $\hat{h}$ is the Fourier transform of h ($\hat{h}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} h(y)e^{-ixy}dy$), $\mathcal{P}$ is now the set of all primitive[4] periodic orbits[5] and ℓ is the geodesic distance corresponding to (5.4).

The similarity between (5.3) and (5.5) may make you wish that prime numbers would correspond to primitive orbits, with lengths $\log p$, $p \in \mathcal{P}$. No result in this direction is known however, and it seems safer not to think about this analogy as a conjecture, but rather just as a tool guiding intuition (as done e.g. in [3] to understand the pair correlations between the zeros of ζ). Nevertheless, the reader could prove that, as a consequence of Selberg's trace formula, the number of primitive orbits with length less than x is

$$|\{\ell(p) < x\}| \underset{x\to\infty}{\sim} \frac{e^x}{x}.$$

Similarly, by the prime number theorem,

$$|\{\log(p) < x\}| \underset{x\to\infty}{\sim} \frac{e^x}{x}.$$

Montgomery's theorem. A more quantitative connection of analytic number theory with a spectral problems appeared in the early 70's thanks to a conversation, during tea time, in Princeton, about some research on the spacings between the ζ zeros. Here is a how the author of this work, Hugh Montgomery, relates this "serendipity" moment [6].

[4] i.e. not the repetition of shorter periodic orbits

[5] of the geodesic flow on $\Gamma\backslash\mathbb{H}$

I took afternoon tea that day in Fuld Hall with Chowla. Freeman Dyson was standing across the room. I had spent the previous year at the Institute and I knew him perfectly well by sight, but I had never spoken to him. Chowla said: " Have you met Dyson? " I said no, I hadn't. He said: "I'll introduce you." I said no, I didn't feel I had to meet Dyson. Chowla insisted, and so I was dragged reluctantly across the room to meet Dyson. He was very polite, and asked me what I was working on. I told him I was working on the differences between the non-trivial zeros of Riemann's zeta function, and that I had developed a conjecture that the distribution function for those differences had integrand $1 - \left(\frac{\sin \pi u}{\pi u}\right)^2$. *He got very excited. He said: "That's the form factor for the pair correlation of eigenvalues of random Hermitian matrices!" I'd never heard the term "pair correlation." It really made the connection. The next day Atle (Selberg) had a note Dyson had written to me giving references to Mehta's book, places I should look, and so on. To this day I've had one conversation with Dyson and one letter from him. It was very fruitful. I suppose by this time the connection would have been made, but it was certainly fortuitous that the connection came so quickly, because then when I wrote the paper for the proceedings of the conference, I was able to use the appropriate terminology and give the references and give the interpretation. I was amused when, a few years later, Dyson published a paper called "Missed Opportunities." I'm sure there are lots of missed opportunities, but this was a counterexample. It was real serendipity that I was able to encounter him at this crucial juncture.*

So what was it exactly that Montgomery proved? To state his result, we need to first introduce some notation. First by choosing for h an appropriate approximation of an indicator function, from the explicit formula (5.3) one can prove the following: the number of ζ zeros ρ counted with multiplicities in $0 < \Im(\rho) < t$ is asymptotically

$$\mathcal{N}(t) \underset{t\to\infty}{\sim} \frac{t}{2\pi} \log t. \tag{5.6}$$

In particular, the mean spacing between ζ zeros at height t is $2\pi/\log t$. Now, we write as previously $1/2 \pm \mathrm{i}\gamma_n$ for the zeta zeros counted with multiplicity, assuming the Riemann hypothesis and the ordering $\gamma_1 \leq \gamma_2 \leq \dots$ Let $\omega_n = \frac{\gamma_n}{2\pi} \log \frac{\gamma_n}{2\pi}$. From (5.6) we know that $\delta_n = \omega_{n+1} - \omega_n$ has a mean value 1 as $n \to \infty$. A more precise understanding of the zeta zeros interactions relies on the study of the spacings distribution function below for $t \to \infty$,

$$\frac{1}{\mathcal{N}(t)} |\{(n,m) \in [\![1, \mathcal{N}(t)]\!]^2 : \alpha < \omega_n - \omega_m < \beta, n \neq m\}|,$$

and more generally on the operator

$$\tilde{r}(f,t) = \frac{1}{\mathcal{N}(t)} \sum_{1 \leq j,k \leq \mathcal{N}(t), j \neq k} f(\omega_j - \omega_k).$$

As we saw in (5.2), if the ω_k's behaved as independent random variables (up to the ordering), $\tilde{r}(f,t)$ would converge to $\int_{\mathbb{R}} f(y)\mathrm{d}y$ as $t \to \infty$. The following result by Montgomery [10] proves that the zeros are actually not asymptotically independent, but present some statistical repulsion instead. We include an outline of a proof directly following the statement for the interested reader.

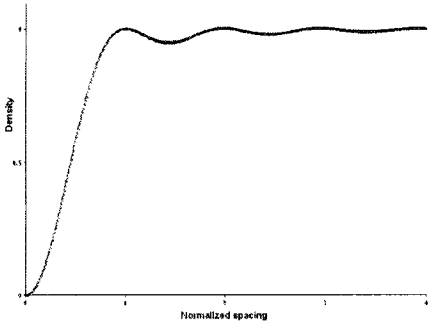


Figure 5.2: The function $\tilde{r}(y)$ and the histogram of the normalized spacing between non-necessarily consecutive ζ zeros, at height 10^{13} (a number of 2×10^9 zeros have been used to compute the empirical density, represented as small circles). Source: Xavier Gourdon [7]

Theorem. Assume the Riemann hypothesis. Suppose f is a test function with the following property: its Fourier transform[6] is $\mathscr{C}^\infty$ and supported in $(-1, 1)$. Then

$$\tilde{r}(f,t) \underset{t\to\infty}{\longrightarrow} \int_{\mathbb{R}} f(y)\tilde{r}(y)\mathrm{d}y,$$

where $\tilde{r}(y) = 1 - \left(\frac{\sin(\pi y)}{\pi y}\right)^2$.

In fact an important conjecture due to Montgomery asserts that the above result holds with no condition on the support of the Fourier transform. However, weakening the restriction even to $\operatorname{supp} \hat{f} \subset (-1-\epsilon, 1+\epsilon)$ for some $\epsilon > 0$ out of reach with known techniques. The Montgomery conjecture would have important consequences for example in terms of the statistics of gaps between the prime numbers $p_1 < p_2 < \ldots$: for example, it would imply that $p_{n+1} - p_n \ll \sqrt{p_n \log p_n}$.

Sketch of proof of Montgomery's Theorem. Consider the function

$$F(\alpha, t) = \frac{1}{\frac{t}{2\pi}\log t} \sum_{0<\gamma,\gamma'<t} t^{\mathrm{i}\alpha(\gamma-\gamma')} \frac{4}{4+(\gamma-\gamma')^2},$$

where the γ's are the imaginary parts of the ζ zeros. This is the Fourier transform of the normalized spacings, up to the factor $4/(4+(\gamma-\gamma')^2)$, present here just for technical convergence reasons. This function naturally appears when counting the second order moments

$$\int_0^t |G(s, t^\alpha)|^2 \mathrm{d}s = F(\alpha, t) t \log t + \mathrm{O}(\log^3 t),\ G(s,x) = 2\sum_\gamma \frac{x^{\mathrm{i}\gamma}}{1+(s-\gamma)^2}. \tag{5.7}$$

As G is a linear functional of the zeros, it can be written as a sum over primes by an appropriate explicit formula like (5.3): Montgomery proved that

$$G(s,x) = -\sqrt{x}\left(\sum_{n\le x} \Lambda(n)\left(\frac{x}{n}\right)^{-\frac{1}{2}+\mathrm{i}s} + \sum_{n>x} \Lambda(n)\left(\frac{x}{n}\right)^{\frac{3}{2}+\mathrm{i}s}\right) + \varepsilon(s,x),$$

[6]Contrary to the Weil and Selberg formulas (5.3) and (5.5), the chosen normalization here is $\hat{f}(x) = \int_{-\infty}^{\infty} f(y)e^{-\mathrm{i}2\pi xy}\mathrm{d}y$

where $\varepsilon(s,x)$ is an error term which, under the Riemann hypothesis, can be bounded efficiently and makes no contribution in the following asymptotics. The moment (5.7) can therefore be expanded as a sum over primes, and the Montgomery-Vaughan inequality (cf. the exercise hereafter) leads to

$$\int_0^t |G(s,t^\alpha)|^2 \mathrm{d}s = (t^{-2\alpha}\log t + \alpha + \mathrm{o}(1))t\log t. \tag{5.8}$$

These asymptotics can be proved by the Montgomery Vaughan inequality, but only in the range $\alpha \in (0,1)$, which explains the support restriction in the hypotheses. Gathering both asymptotic expressions for the second moment of G yields $F(\alpha,t) = t^{-2\alpha}\log t + \alpha + \mathrm{o}(1)$. Finally, by the Fourier inversion formula,

$$\frac{1}{\frac{t}{2\pi}\log t}\sum_{0\le\gamma,\gamma'\le t} f\left((\gamma-\gamma')\frac{\log t}{2\pi}\right)\frac{4}{4+(\gamma-\gamma')^2} = \int_{\mathbb{R}} F(\alpha,t)\hat{f}(\alpha)\mathrm{d}\alpha.$$

If $\mathrm{supp}\,\hat{f} \subset (-1,1)$, this is approximately

$$\int_{\mathbb{R}} \hat{f}(\alpha)(t^{-2|\alpha|}+|\alpha|)\mathrm{d}\alpha = \int_{\mathbb{R}} e^{-2|\alpha|}\hat{f}(\alpha/\log t)\mathrm{d}\alpha + \int_{\mathbb{R}} |\alpha|\hat{f}(\alpha)\mathrm{d}\alpha$$

$$= \hat{f}(0)+f(0)-\int_{\mathbb{R}}(1-|\alpha|)\hat{f}(\alpha)\mathrm{d}\alpha+\mathrm{o}(1) = f(0)+\int_{\mathbb{R}} f(x)\left(1-\left(\frac{\sin\pi x}{\pi x}\right)^2\right)\mathrm{d}x+\mathrm{o}(1),$$

by the Plancherel formula. □

(Difficult) Exercise. Let (a_r) be complex numbers, (λ_r) distinct real numbers and

$$\delta_r = \min_{s\neq r}|\lambda_r-\lambda_s|.$$

Then the Montgomery-Vaughan inequality asserts that

$$\frac{1}{t}\int_0^t |\sum_r a_r e^{\mathrm{i}\lambda_r s}|^2\mathrm{d}s = \sum_r |a_r|^2\left(1+\frac{3\pi\theta}{t\delta_r}\right)$$

for some $|\theta|<1$. In particular,

$$\int_0^t \left|\sum_{n=1}^\infty \frac{a_n}{n^{\mathrm{i}s}}\right|^2 \mathrm{d}s = t\sum_{n=1}^\infty |a_n|^2 + \mathrm{O}\left(\sum_{n=1}^\infty n|a_n|^2\right).$$

Prove that the above result implies (5.8).

To numerically test Montgomery's conjecture, Odlyzko [11] computed the normalized gaps, $\omega_{i+1}-\omega_i$, and produced the joint histogram. In particular, note that the limiting density vanishes at 0, contrasting with Figure 1, and that this type of repulsion coincides remarkably with the shape of gaps for random matrices.

Moreover, Montgomery's result has been extended in the work by Rudnick and Sarnak [14], who proved that for some statistics depending on more than just one gap, the ζ zeros also present the same limit distribution as predicted by Random Matrix Theory. This urges us to explain in more details what we mean by *random matrices*.

5.3 Eigenvalues repulsion for random matrices

Let χ be a point process, i.e. a random set of points $\{x_1,x_2,\dots\}$, in a metric space Λ, identified with the random punctual measure $\sum_i \delta_{x_i}$. The kth correlation function for this point process, ρ_k,

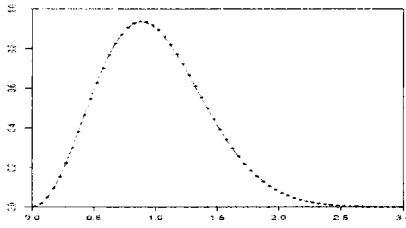


Figure 5.3: The distribution function of asymptotic gaps between eigenvalues of random matrices compared with the histogram of gaps between successive normalized ζ zeros, based on a billion zeros near $\#1.3 \cdot 10^{16}$.

is defined as the asymptotic (normalized) probability of having exactly one particle in respective neighborhoods of k fixed points. More precisely, if the u_i's are distinct in Λ,

$$\rho_k(u_1, \ldots, u_k) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(\chi(B_{u_i,\epsilon}) = 1, 1 \leq i \leq k)}{\prod_{j=1}^{k} \lambda(B_{u_i,\epsilon})},$$

provided that the limit exists (here $B_{u_i,\epsilon}$ denotes the ball with radius ϵ and center u, and the measure λ will be specified later). If χ consists almost surely of n points, the correlation functions satisfy the integration property

$$(n-k)\rho_k(u_1, \ldots, u_k) = \int_{\Lambda} \rho_{k+1}(u_1, \ldots, u_{k+1}) \mathrm{d}\lambda(u_{k+1}). \tag{5.9}$$

Interestingly, many properties about a point process are well-understood when the correlation functions are also determinants. More precisely, assume now that $\Lambda = \mathbb{C}$. If there exists a function $K : \mathbb{C} \times \mathbb{C} \to \mathbb{C}$ such that for all $k \geq 1$ and $(z_1, \ldots, z_k) \in \mathbb{C}^k$

$$\rho_k(z_1, \ldots, z_k) = \det\left(K(z_i, z_j)_{i,j=1}^{k}\right),$$

then χ is said to be a determinantal point process with respect to the underlying measure λ and with correlation kernel K.

The determinantal condition for *all* correlation functions is quite restrictive. Nevertheless, as stated in the following theorem, any bidimensional system of particles with quadratic interaction is determinantal (see [1] for a proof).

Theorem. Let $\mathrm{d}\lambda$ be any[7] finite measure on $\mathbb{C}$ (eventually concentrated on a line). Consider the probability distribution with density

$$c(n) \prod_{1 \leq k < l \leq n} |z_l - z_k|^2$$

with respect to $\prod_{j=1}^{n} \mathrm{d}\lambda(z_j)$, where $c(n)$ is the normalization constant. For this joint distribution, $\{z_1, \ldots, z_n\}$ is a determinantal point process with the following explicit kernel,

$$K(x, y) = \sum_{k=0}^{n-1} P_k(x)\overline{P_k(y)}$$

where P_k $(0 \leq k \leq n-1)$ is a polynomial with degree k and the P_j's are orthonormal for the Hermitian product $f, g \mapsto \int f\bar{g}\mathrm{d}\lambda$.

[7] We just need a decreasing of the mass at infinity of type $\int_{|z|>t} \mathrm{d}\lambda(z) \ll t^{-k}$ for any $k > 0$.

We apply the above result to the following examples, which are among the most studied random matrices. First, consider the so-called Gaussian unitary ensemble (GUE). This is the ensemble (or set) of random $n \times n$ Hermitian matrices with independent (up to symmetry) Gaussian entries: $M_{ij}^{(n)} = \overline{M_{ji}^{(n)}} = \frac{1}{\sqrt{n}}(X_{ij} + \mathrm{i}Y_{ij}), 1 \leq i < j \leq n$, where the X_{ij}'s and Y_{ij}'s are independent centered real Gaussians entries with mean 0 and variance $1/2$ and $M_{ii}^{(n)} = X_{ii}/\sqrt{n}$ with X_{ii} real centered Gaussians with variance 1, still independent. These random matrices are natural in the sense that they are uniquely characterized by the independence (up to symmetry) of their entries, and invariance by unitary conjugacy. A similar natural set of matrices, when the entries are now real Gaussian, called GOE (Gaussian orthogonal ensemble) will appear in the next section.

For the GUE, the distribution of the eigenvalues has an explicit density,

$$\frac{1}{Z_n} e^{-n\sum_{i=1}^n \lambda_i^2/2} \prod_{1\leq i<j\leq n} |\lambda_i - \lambda_j|^2 \tag{5.10}$$

with respect to Lebesgue measure (see e.g. [1] for a derivation of this result). We denote by (h_n) the Hermite polynomials, more precisely the successive monic polynomials orthogonal with respect to the Gaussian weight $e^{-x^2/2}\mathrm{d}x$, and consider the associated normalized functions

$$\psi_k(x) = \frac{e^{-x^2/4}}{\sqrt{\sqrt{2\pi}k!}} h_k(x).$$

Then from the previous Theorem, one can prove that the set of point $\{\lambda_1, \ldots, \lambda_n\}$ with law (5.10) is a determinantal point process whose kernel (with respect to the Lebesgue measure on $\mathbb{R}$) is given by

$$K^{\mathrm{GUE}(n)}(x,y) = n\frac{\psi_n(x\sqrt{n})\psi_{n-1}(y\sqrt{n}) - \psi_{n-1}(x\sqrt{n})\psi_n(y\sqrt{n})}{x-y},$$

extended by continuity when $x = y$. Here we used a simplification: the sum over all orthogonal polynomials can simplify as a sum over just two of them, this is the Christoffel-Darboux formula.

The Plancherel-Rotach asymptotics for the Hermite polynomials implies that, as $n \to \infty$, $K^{\mathrm{GUE}(n)}(x,x)/n$ has a non-trivial limit.

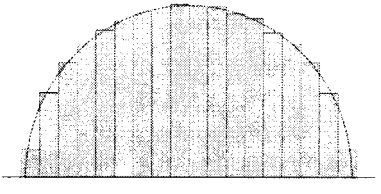


Figure 5.4: Histogram of the eigenvalues from the Gaussian Unitary Ensemble in dimension 10^4. Dashed: the rescaled semicircle law.

More precisely, the empirical spectral distribution $\frac{1}{n}\sum \delta_{\lambda_i}$ converges in probability to the semicircle law with density

$$\rho_{sc}(x) = \frac{1}{2\pi}\sqrt{(4-x^2)_+}$$

with respect to Lebesgue measure. This is the asymptotic behavior of the spectrum in the macroscopic regime. The microscopic interactions between eigenvalues also can be evaluated thanks to

asymptotics of the Hermite orthogonal polynomials: for any $x \in (-2, 2)$, $u \in \mathbb{R}$,

$$\frac{1}{n\rho_{sc}(x)} K^{\mathrm{GUE}(n)}\left(x, x + \frac{u}{n\rho_{sc}(x)}\right) \underset{n\to\infty}{\longrightarrow} K(u) = \frac{\sin(\pi u)}{\pi u}.$$

This leads to a repulsive correlation structure for the eigenvalues at the scale of the average gap: for example the two-point correlation function asymptotics are

$$\left(\frac{1}{n\rho_{sc}(x)}\right)^2 \rho_2^{\mathrm{GUE}(n)}\left(x, x + \frac{u}{n\rho_{sc}(x)}\right) \underset{n\to\infty}{\longrightarrow} \tilde{r}(u) = 1 - \left(\frac{\sin(\pi u)}{\pi u}\right)^2,$$

the strict analogue to Montogmery's result, an analogy identified by Dyson as mentioned in Section 2.

[h]

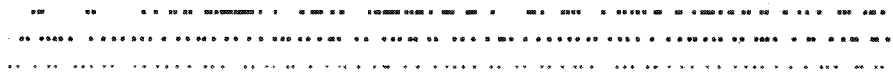


Figure 5.5: Upper line: a sample of independent points distributed according to the semicircle law after zooming in the bulk. Middle line: a sample eigenvalues of the GUE after zooming in the bulk of the spectrum. Lower line: a sequence of imaginary parts of the ζ zeros, about height 10^5.

A remarkable fact about the above limiting sine kernel is that it appears universally in the limiting correlation functions of random Hermitian matrices with independent (up to symmetry) entries (not necessarily Gaussian); these deep universality results were achieved, still for the Hermitian symmetry class, in recent works by Erdős, Yau et al, or by Tao, Vu. In the case of other symmetry classes[8], the universality of the local eigenvalues statistics has also been proved by Erdős, Yau et al.

Finally we want to mention the following *structural* reason for the repulsion of the eigenvalues of typical matrices: as an exercise, the reader could prove that the space of Hermitian matrices with at least one repeated eigenvalue has codimension 3 in the space of all Hermitian matrices. Repeated eigenvalues therefore occur with very small probability compared to independent points (on a product space, the codimension of the subspace where two points coincide is 1). László Erdős asked me about a *structural*, heuristic, argument for the repulsion of the ζ zeros. Unable to answer it, I transmit the question to the readers.

5.4 Eigenvalues repulsion for quantum billiards

To conclude this expository note, we wish to mention some conjectures about the asymptotic distribution of eigenvalues, for the Laplacian on compact spaces.

The examples we consider are two-dimensional quantum billiards[9]. For some billiards, the classical trajectories are integrable[10] and for others they are chaotic.

On the quantum side, we consider the Helmholtz equation inside the billiard, describing the standing waves:

$$-\Delta\psi_n = \lambda_n \psi_n,$$

where the spectrum is discrete as the domain is compact, with ordered eigenvalues $0 \leq \lambda_1 \leq \lambda_2 \ldots$, and appropriate Dirichlet or Neumann boundary conditions. The questions about quantum

[8] i.e. for random symmetric matrices or random symplectic matrices

[9] A billiard is a compact connected set with nonempty interior, with a generally piecewise regular boundary, so that the classical trajectories are straight lines reflecting with equal angles of incidence and reflection

[10] Roughly speaking this means that there are many conserved quantities along the trajectory, and that explicit solutions can be given for the speed and position of the ball at any time

[h]

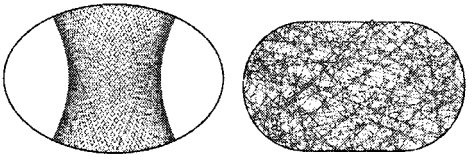


Figure 5.6: An integrable billiard (ellipse) and a chaotic one (stadium)

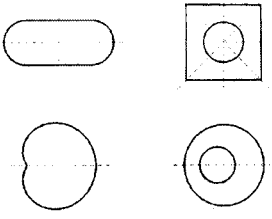


Figure 5.7: Some chaotic billiards, from left to right, up to down: the stadium, Sinai's billard, the cardioid, and a billiard with no name.

billiards we are interested here is about the asymptotic behavior of the λ_n's, i.e. whether they will present asymptotic independence or a Random Matrix Theory type of repulsion. The situation is still somehow mysterious: there is a conjectural dichotomy between the chaotic and integrable cases.

First, in 1977, Berry and Tabor [4] put forward the conjecture that for most integrable systems, the large eigenvalues have the statistics of a Poisson point process, i.e. rescaled gaps being asymptotically exponential random variables, like in Section 1. More precisely, by Weyl's law, we know that the number of such eigenvalues up to λ is

$$|\{i : \lambda_i \leq \lambda\}| \underset{\lambda \to \infty}{\sim} \frac{\text{area}(\mathcal{D})}{4\pi} \lambda. \tag{5.11}$$

To analyze the correlations between eigenvalues, consider the point process

$$\chi^{(n)} = \frac{1}{n} \sum_{i \leq n} \delta_{\frac{4\pi}{\text{area}(\mathcal{D})}(\lambda_{i+1} - \lambda_i)}.$$

Its expectation converges to 1 (as $n \to \infty$) from (5.11). By the conjectured limiting Poissonian behavior, the spacing distribution converges to an exponential law: for any $I \subset \mathbb{R}^+$

$$\chi^{(n)}(I) \underset{n \to \infty}{\longrightarrow} \int_I e^{-x} \mathrm{d}x. \tag{5.12}$$

In the chaotic case, the situation differs radically: the eigenvalues are supposed to repel each other, with gaps statistics conjecturally similar to those of a random matrix, from an ensemble depending on the symmetry properties of the system (e.g. time-reversibility for our quantum billiards correspond to the Gaussian Orthogonal Ensemble). This is known as the Bohigas-Giannoni-Schmidt Conjecture [5].

Numerical experiments were performed in [5] giving a correspondence between the eigenvalue spacings statistics for Sinai's billiard and those of the Gaussian Orthogonal Ensemble. The

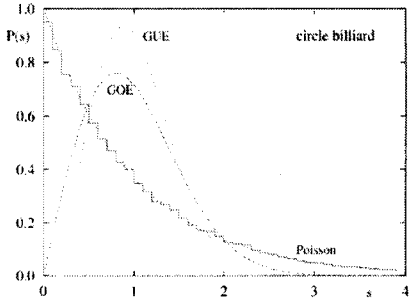


Figure 5.8: Energy levels for the circular billiard compared to those of the Gaussian ensembles and Poissonian statistics (data and picture from [2]).

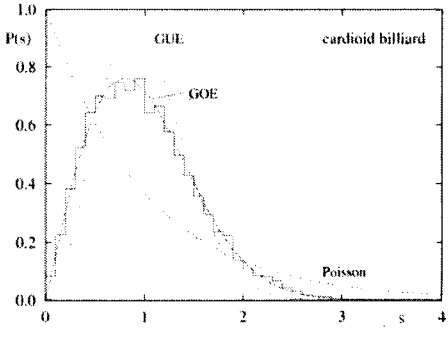


Figure 5.9: Energy levels for the cardioid billiard compared to those of the Gaussian ensembles and Poissonian statistics (data and picture from [2]).

joint graphs, by A. Backer, present similar experiments for an integrable billiard (Figure 8) and a chaotic one (Figure 9). These statistics are perfectly coherent with both the Berry-Tabor and the Bohigas-Giannoni-Schmidt conjectures. This deepens the interest in these Random Matrix Theory distributions, which appear increasingly in many fields, including analytic number theory.

References

[1] G.W. Anderson, A. Guionnet, O. Zeitouni, *An Introduction to Random Matrices*, Cambridge University Press, 2009.

[2] A. Backer, Ph.D. thesis, Universitat Ulm, Germany, 1998.

[3] M. V. Berry, J. P. Keating, The Riemann zeros and eigenvalue asymptotics, SIAM Review 41 (1999), 236–266.

[4] M.V. Berry, M. Tabor, *Level clustering in the regular spectrum*, Proc. Roy. Soc. Lond. A **356** (1977), 375–394.

[5] O. Bohigas, M.-J. Giannoni, C. Schmidt, *Characterization of chaoic quantum spectra and universality of level fluctuation laws*, Phys. Rev. Lett. **52** (1984), 1–4.

[6] J. Derbyshire, Prime Obsession: Bernhard Riemann and the Greatest Unsolved Problem in Mathematics (Plume Books, 2003)

[7] X. Gourdon, The 10^{13} first zeros of the Riemann Zeta function, and zeros computation at very large height.

[8] J.P. Keating, N.C. Snaith, Random matrix theory and number theory, in The Handbook on Random Matrix Theory, 491–509, edited by G. Akemann, J. Baik& P. Di Francesco, Oxford university Press, 2011.

[9] M. L. Mehta, *Random matrices*, Third edition, Pure and Applied Mathematics Series **142**, Elsevier, London, 2004.

[10] H.L. Montgomery, *The pair correlation of zeros of the zeta function*, Analytic number theory (Proceedings of Symposium in Pure Mathemathics **24** (St. Louis Univ., St. Louis, Mo., 1972), American Mathematical Society (Providence, R.I., 1973), pp. 181–193.

[11] A.M. Odlyzko, *On the distribution of spacings between the zeros of the zeta function*, Math. Comp. **48** (1987), 273–308.

[12] N.M. Katz, P. Sarnak, *Random Matrices*, Frobenius Eigenvalues and monodromy, American Mathematical Society Colloquium Publications, 45. American Mathematical Society, Providence, Rhode island, 1999.

[13] B. Riemann, *Über die Anzahl der Primzahlen unter einer gegebenen Grösse*, Monatsberichte der Berliner Akademie, Gesammelte Werke, Teubner, Leipzig, 1892.

[14] Z. Rudnick, P. Sarnak, *Zeroes of principal L-functions and random matrix theory*, Duke Math. J. **81** (1996), no. 2, 269–322. A celebration of John F. Nash.

[15] E. C. Titchmarsh, The Theory of the Riemann Zeta Function, London, Oxford University Press, 1951.

FEATURE

6

MATHEMATICAL MINUTIAE

Four Proofs of the Integer Side Theorem

Evan O'Dorney[†]
Harvard University '15
Cambridge, MA 02138
odorney@college.harvard.edu

Consider the following theorem:

Theorem 1. *Suppose that a rectangle R is tiled with a finite number of rectangular tiles $T_1, \ldots, T_n$ and that each tile T_i has at least one integer side. Then the large rectangle R also has at least one integer side.*

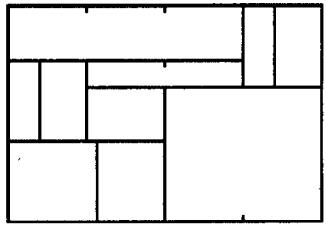


Figure 6.1: An illustration of the Integer Side Theorem. Tick marks show that the larger rectangles have side-lengths divisible by the unit.

The origins of this theorem are somewhat murky. As early as 1903, Dehn discovered, in a serendipitous way, the analogous statement for rational sides; the integer version can be traced to a result of de Bruijn in 1969 [1]. The formulation given above (which differs from de Bruijn's, but seems to have become something of a standard combinatorial problem) will henceforth be referred to as the Integer Side Theorem.

At first glance, the conditions of the theorem may seem unmotivated; however, it can be used to solve several tiling puzzles, for instance:

Problem. *Can a 10×10 square board be covered with 1×4 rectangular tiles?*

Solution. *If we scale down the square board a factor of 4, we are asking for a tiling of a 2.5×2.5 board by 1×0.25 tiles, which we see is not possible using the Integer Side Theorem.*

While the applications of the Integer Side Theorem tend to remain firmly in the realm of tiling, the proofs do not. There are many known proofs, which use methods ranging from elementary combinatorics to seemingly unrelated ideas in analysis and number theory. In the following pages, I will aim to present them in their great variety.

[†]Evan O'Dorney is a Harvard freshman planning to concentrate in mathematics with a possible secondary field in music. He has been homeschooled up through high school and is most famous for winning the Scripps National Spelling Bee in 2007. He enjoys a wide range of mathematical topics, especially number theory. His nonmathematical interests include juggling, calculator programming, and improvising classical piano music.

It will be convenient to talk about the coordinates of various points in the tiling. For this purpose, we let R lie with two adjacent sides on the positive x and y axes, meeting at a vertex at $(0,0)$.

Integration. The first proofs found were based on calculus. The one presented here has the additional twist of using calculus with complex numbers!

Consider the result of integrating

$$f(x,y) = e^{2\pi i(x+y)}$$

over a rectangle $[a,b] \times [c,d] \subset \mathbb{R}^2$. By separating the integral into

$$\int_a^b e^{2\pi ix}dx \cdot \int_c^d e^{2\pi iy}dy,$$

we find that it is not hard to integrate by elementary calculus and it equals

$$\frac{e^{2\pi ib} - e^{2\pi ia}}{2\pi i} \cdot \frac{e^{2\pi id} - e^{2\pi ic}}{2\pi i}.$$

This value is 0 if and only if

$$e^{2\pi ia} = e^{2\pi ib} \quad \text{or} \quad e^{2\pi ic} = e^{2\pi id},$$

that is, if and only if one of the sides $b-a$ and $d-c$ is an integer.

The jig is up. If each of the tiles T_i in a tiling has an integer side, then f integrates to 0 on each of them. Then f integrates to 0 on the entire rectangle R, which must therefore have an integer side. □

Checkerboard. Tiling problems are often tackled by coloring. In fact, Problem 6 above regarding the 10×10 square has a very concise solution based on dividing the board into 2×2 blocks and coloring them black and white in checkerboard fashion. If we scale down by a factor of 4 as in the solution above, we are led to considering a checkerboard with squares of side $1/2$. This is the inspiration for our next proof, which has at least two independent discoverers (Richard Rochberg and Sherman Stein).

We draw a checkerboard with squares of side $1/2$; one of them, say a black one, is fixed in the first quadrant with its lower left corner at the origin. If a rectangle with one side length 1 is placed anywhere on this checkerboard with its sides parallel to the axes, then it can be cut into two strips of width $1/2$, whose colors are inverses of each other (see Figure 6). This shows that the rectangle has equal areas of black and white. The same conclusion can be proved for an $n \times x$ rectangle, for n an integer, simply by cutting it into $1 \times x$ rectangles.

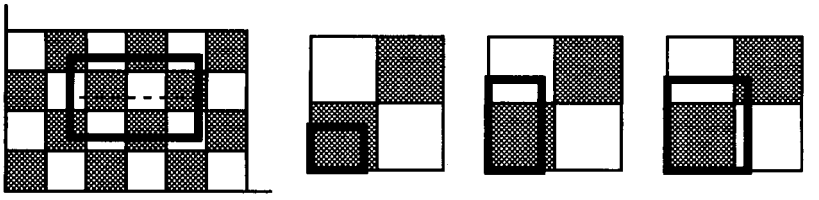


Figure 6.2: A $1 \times x$ rectangle on a checkerboard of $1/2 \times 1/2$ squares has equal black and white areas; an $a \times b$ rectangle, where $a, b < 1$ and one vertex is at $(0,0)$, does not.

Thus, in our given tiling, each of the component tiles T_i has equal areas of black and white. So, it remains to derive a contradiction by showing that the large rectangle R has unequal black and white areas if both its sides, a and b, are non-integers.

If $a \geq 1$, we can clearly diminish a by 1 repeatedly, without changing the difference between the black and white areas. After doing the same with b, we are left with a rectangle that fits inside a 1×1 square and overlaps either one, two, or four of the cells of the checkerboard, as shown in Figure 6. The reader is invited to derive, in each of the three cases, formulas for the black and white areas and see that, indeed, the black area is always larger. □

Counting. Our third proof uses a technique less associated with tiling and more with combinatorics in general: counting a set of objects in two ways. This is one of a family of purely combinatorial proofs that generalizes readily to the analogous theorems not just about *integer* sides but also about *rational* sides, *algebraic* sides, or any given additive subgroup of the reals.

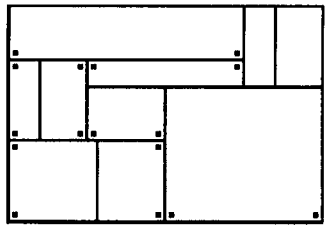


Figure 6.3: Dots mark vertices with integer coordinates.

In each rectangle T_i of the tiling, we place a dot next to each vertex of the rectangle that has both coordinates integers. Thus we are marking pairs (T, V), where T is a tile, V is one of its four vertices, and V has both coordinates integers. It is clear that two vertices of the same tile which are connected by one of its integer-length sides are either both dotted or both undotted; this implies that the number of dots on the tile—and hence on the whole tiling—is even.

On the other hand, each vertex *except* for the four corners of the large rectangle R has either two or four rectangles of which it is a vertex, giving two or four dots (assuming the vertex in fact has integer coordinates). Thus the total number of dots at these vertices is even. We deduce that the total number of dots at the corners of R is even. Note that R automatically has one dot (at the origin). Since there must be another, we deduce that R has an integer side. □

Primes. Finally, we present an amusing argument by Raphael Robinson, in which the infinitude of primes (which is itself a theorem with a wide assortment of proofs) enters in an essential way.

Pick a prime p, and scale the entire tiling up by a factor of p. Then move each vertex (x, y) of the resulting tiling to the next lower lattice point $(\lfloor x \rfloor, \lfloor y \rfloor)$. This may reduce some tiles to degeneracy, but it will not destroy their rectangularity or ability to tile the newly enlarged large rectangle R'.

Now, of course, all tiles have integer sides, but those which originally had an integer side now have a side divisible by p and hence area divisible by p. It follows that R' has area divisible by p and—since p is prime—either its height or its width is divisible by p.

We now return to the original rectangle R. At least one of its dimensions must be within $1/p$ of an integer, and this conclusion holds for every prime p. So one of its two dimensions is within $1/p$ of an integer for *infinitely* many primes p and therefore is an integer. □

These four proofs demonstrate how a simple problem in tiling can bring together ideas from many different areas of mathematics. The third proof is notable for sticking to the concepts of rectangles and integers that appear in the problem; the foreign objects that appear in the others—complex exponentials, colored squares of side $1/2$, and dilation by prime ratios—mark them as examples of a more imaginative style of thinking.

Generalizations of the Integer Side Theorem, which apply to higher dimensions and to tilings of cylinders, tori, and other exotic spaces, have also been found. The theorem also has an interesting "converse" of sorts, which is seldom if ever mentioned:

Theorem 2. *Suppose that a rectangle R is tiled with a finite number of rectangular tiles $T_1, \ldots, T_n$. Suppose that R and the tiles T_1 through T_{n-1} each have an integer side. Then T_n also has at least one integer side.*

The reader is invited to investigate how many of the four proofs above can be adapted to this statement.

References

[1] S. Wagon, *Fourteen proofs of a result about tiling a rectangle*, Amer. Math. Monthly, Vol. 94 (1987), 601–617 (available at `http://mathdl.maa.org/images/upload_library/22/Ford/Wagon601-617.pdf`)

FEATURE

7

APPLIED MATHEMATICS CORNER

Minimum Variance Inflation Hedge

Adam Arthurs[†]
Harvard University '12
Cambridge, MA 02138
arthurs@fas.harvard.edu

Abstract

By defining the appropriate pre-Hilbert space, this paper proves that, given a fixed set of assets, known expectations and pairwise covariances of the returns thereof, and a fixed overall return, there always exists a unique portfolio of these assets that is expected to achieve this overall return with minimum variance. This is a well-known result in portfolio theory but the proof given here is novel. Using historical data from 1940 through 2011, this result is applied to construct the minimum variance portfolio for 2012 that is expected to return the rate of inflation. In theory, this portfolio is the least risky way to protect against inflation during 2012; in practice, this is likely not the case. Backtests are performed to show that such minimum variance portfolios don't perform as expected in any single year, but do on average over several years, potentially leading to useful longer-term investment strategies.

7.1 The Problem

Suppose, and realistically so, that an investor knows roughly the rate of inflation that will prevail over the next 365 days. Call it r on an annualized basis. This means that \$1 in one year is worth only $\$1/(1+r)$ today. In order to protect against the devaluation of his cash, the investor demands an annual return of r annually. In the past, he could achieve this simply by putting his money in a bank account with interest rate r. However, the projected inflation rate for 2012 is $r \approx .02$ whereas one-year bank accounts[1] opened today have an interest rate of only roughly .01.

Bank accounts are zero-variance[2]: the depositor locks in a known, fixed, annual interest rate when he deposits his money and he necessarily gets back his deposit plus the appropriate amount of interest in one year. The rate of return is a constant, not a random variable. In order to get annual returns greater than .01, which is what the investor demands to combat inflation ($r \approx .02$), it is necessary to invest in nonzero-variance assets. Greater returns usually necessitate greater risk (greater variance).

The investor seeks to combat inflation while minimizing the variance of his portfolio: he wants the lowest risk inflation hedge. More generally, the investor demands an annual return of r and seeks to minimize the variance of his portfolio. How much of which assets should he invest in for arbitrary r?

[†] Adam Arthurs is a senior Applied Math for Physics concentrator who lives in Adams house and has been a course assistant for Math 23, 110, 116, and 117. Apart from academics, Adam is the catcher for the club baseball team and a decidedly mediocre shooting guard for the B-league basketball intramural champions. Adam will go to work for a hedge fund upon graduation.

Since completing this paper, Adam used the fixed-return minimum-variance result to create an investment strategy which he implemented at the end of February; as of the beginning of April, it has returned 13%.

[1] Certificate of Deposit (CD; the simplest way to put money in the bank) rates can be looked up easily online.

[2] Ignoring the risk of bank default.

7.2 Norms and inner products

In a normed vector space V, the *norm* of any vector $x \in V$, denoted $||x||$, is a generalization of distance or length. For example, $\mathbb{R}^3$ has the well-known Euclidean norm $||(x_1, x_2, x_3)|| = \left({x_1}^2 + {x_2}^2 + {x_3}^2\right)^{1/2}$. The following properties are necessary and sufficient for defining a norm:

1. For all $x \in V$, $||x|| \geq 0$, and $||x|| = 0$ only if x is the zero vector,
2. For all $\alpha \in \mathbb{R}, x \in V$, $||\alpha x|| = |\alpha| \cdot ||x||$,
3. For all $x, y \in V$, $||x + y|| \leq ||x|| + ||y||$.

Note that the specific norm used in any given space is a property of the space itself. For example, $\mathbb{R}^3$ is the space of all 3-component vectors of real numbers with the Euclidean norm. One could also define another space that is the space of all 3-component vectors of real numbers with the norm $||(x_1, x_2, x_3)|| = \left({x_1}^5 + {x_2}^5 + {x_3}^5\right)^{1/5}$ (which is, indeed, a valid norm according to the properties above).

Along with a norm, a vector space V may have an *inner product* defined on it. As with the norm, the specific inner product is an intrinsic property of the space itself. If there is an inner product defined, it takes any two vectors in V and returns a scalar: any inner product is a function from $V \times V$ to $\mathbb{R}$. The inner product of x and y is denoted $(x \mid y)$. The following properties are necessary and sufficient for defining an inner product on any vector space for which the underlying field is the reals[3]:

1. For all $x, y \in V$, $(x \mid y) = (y \mid x)$,
2. For all $x, y, z \in V$, $(x + y \mid z) = (x \mid z) + (y \mid z)$,
3. For all $\alpha \in \mathbb{R}, x, y \in V$, $(\alpha x \mid y) = \alpha(x \mid y)$ for all $\alpha \in \mathbb{R}$, $x, y \in V$,
4. $(x \mid x) \geq 0$ for all $x \in V$ and $(x \mid x) = 0$ only if x is the zero vector.

We will say that a vector space endowed with an inner product as above is a *pre-Hilbert space*.

7.3 Induced norms

We will now show that an inner product on a vector space V induces a norm.

Theorem 1. *For any vector space V with an inner product, and for any $x, y \in V$,*

$$|(x \mid y)|^2 \leq (x \mid x) \cdot (y \mid y)\,. \tag{7.1}$$

Proof. If y is the zero vector, the inequality holds trivially as the equality $0 = 0$. Otherwise, note that the properties of an inner product imply that for all $\alpha \in \mathbb{R}$ and $x, y \in V$,

$$0 \leq (x - \alpha y \mid x - \alpha y) = (x \mid x) - 2\alpha(x \mid y) + \alpha^2(y \mid y).$$

Substituting $\alpha = (x \mid y)/(y \mid y)$ into the previous expression gives $0 \leq (x \mid x) - (x \mid y)^2/(y \mid y)$ which can be rearranged into (7.1), the desired result, using (7.2). □

An important consequence is the following:

Theorem 2. *For any valid inner product, the following function is a valid norm:*

$$||x|| = \sqrt{(x \mid x)}\,. \tag{7.2}$$

Proof. The numbers below correspond to the norm properties from section 2.

[3] If the underlying field is the complex numbers as opposed to the reals, the first property changes to $(x \mid y) = \overline{(y \mid x)}$ and the other properties are unchanged.

1. Follows trivially from inner product property 4.

2. Follows from inner product property 3: $||\alpha x|| = \sqrt{(\alpha x \,|\, \alpha x)} = \sqrt{\alpha^2 (x \,|\, x)} = |\alpha| \cdot ||x||$.

3. Follows from (7.1) and inner product properties 1 and 2: $||x+y||^2 = (x+y \,|\, x+y) = (x \,|\, x) + 2(x \,|\, y) + (y \,|\, y) \leq ||x||^2 + 2||x||\,||y|| + ||y||^2 = (||x|| + ||y||)^2 \;\Rightarrow\; ||x+y|| \leq ||x|| + ||y||$.

□

7.4 Projections

We now show the existence of orthogonal projections in a finite-dimensional real pre-Hilbert space.

Theorem 3. *For any finite-dimensional real pre-Hilbert space H, subspace $M \subset H$, and vector $x \in H$, there exists $m_0 \in M$ such that $||x - m_0|| \leq ||x - m||$ for all $m \in M$, and this m_0 is unique. In other words, in a finite-dimensional pre-Hilbert space, for any subspace and vector x, there is a unique vector in the subspace that is closest to x.*

Proof. For all $y, z \in H$, if $(y \,|\, z) = 0$ then

$$||y+z||^2 = (y+z \,|\, y+z) = ||y||^2 + 2(y \,|\, z) + ||z||^2 = ||y||^2 + ||z||^2 \,. \tag{7.3}$$

Now, using the notation presented in the theorem statement above, suppose that there exists $m_0 \in M$ such that $(x - m_0 \,|\, m) = 0$ for all $m \in M$. Then, for all $m \in M$,

$$||x - m||^2 = ||x - m_0 + m_0 - m||^2 = ||x - m_0||^2 + ||m_0 - m||^2 \tag{7.4}$$

where the second equality is from (7.3), which can be invoked because $m_0 - m \in M$ (M is a subspace) and so $(x - m_0 \,|\, m_0 - m) = 0$. Noting that $||m_0 - m|| > 0$ if $m \neq m_0$ and $||m_0 - m|| = 0$ if $m = m_0$, it holds that $||x - m|| > ||x - m_0||$ for all $m \in M$ for which $m \neq m_0$, and thus m_0 is the unique vector in M that is closest to x. In other words, if there exists $m_0 \in M$ such that $(x - m_0)$ is orthogonal to every vector in M, then this m_0 is the unique vector in M that is closest to x. Thus, if we have a fail-safe method for constructing such a m_0 then the finite-dimensional pre-Hilbert space projection theorem is proved.

H is finite-dimensional and thus any subspace $M \subset H$ must also be finite dimensional. So M must have a finite number n of basis vectors $\{y_1, ..., y_n\}$. By definition of a basis, every vector in M can be expressed as a linear combination of these basis vectors, and thus $(x - m_0)$ is orthogonal to every vector in M if and only if it is orthogonal to each of $\{y_1, ..., y_n\}$. So, $(x - m_0 \,|\, y_i) = 0$ or

$$(m_0 \,|\, y_i) = (x \,|\, y_i) \tag{7.5}$$

for all $i \in \{1, ..., n\}$. But $m_0 \in M$ so it, too, can necessarily be expressed as a linear combination of the basis vectors: there exist $\alpha_1, ..., \alpha_n \in \mathbb{R}$ such that

$$m_0 = \sum_{j=1}^{n} \alpha_j \, y_j \,. \tag{7.6}$$

Substituting (7.6) into the left side of (7.5) for any fixed $i \in \{1, ..., n\}$ gives

$$\sum_{j=1}^{n} \alpha_j (y_j \,|\, y_i) = (x \,|\, y_i) \,. \tag{7.7}$$

The set of equations (7.7) for each $i \in \{1, ..., n\}$ can be written as

$$\underbrace{\begin{bmatrix} (y_1 \mid y_1) & (y_1 \mid y_2) & \dots & (y_1 \mid y_n) \\ (y_1 \mid y_2) & (y_2 \mid y_2) & \dots & (y_2 \mid y_n) \\ \vdots & \vdots & & \vdots \\ (y_1 \mid y_n) & (y_2 \mid y_n) & \dots & (y_n \mid y_n) \end{bmatrix}}_{G} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} (x \mid y_1) \\ (x \mid y_2) \\ \vdots \\ (x \mid y_n) \end{bmatrix} \tag{7.8}$$

where the matrix on the left is known as the Gram matrix and is denoted G. For any basis $\{y_1, ..., y_n\}$ for M, it is possible to perform the Gram-Schmidt procedure on these basis vectors in order to create an orthonormal basis $\{e_1, ..., e_n\}$ for M such that

$$(e_i \mid e_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{7.9}$$

for all $i, j \in \{1, ..., n\}$. Substituting $\{e_1, ..., e_n\}$ into G instead of $\{y_1, ..., y_n\}$ gives the identity matrix I which is invertible. Because it is always possible to use Gram-Schmidt to choose an orthonormal basis for M, it is true that G is always invertible.

Because G is invertible then there is necessarily a solution for $\{\alpha_1, ..., \alpha_n\}$, and via (7.6) we have a fail-safe method of constructing an $m_0 \in M$ such that $(x - m_0)$ is orthogonal to every vector in M. Thus, this m_0 is the unique vector in M that is closest to x and the finite-dimensional pre-Hilbert space projection theorem is proved. □

Theorem 4. *Given a finite-dimensional real pre-Hilbert space H, a subspace $M \subset H$, and a vector $x \in H$, construct the linear variety $N \subset H$ that is M translated by x:*

$$N = M + \{x\} = \left\{ m + x \;\middle|\; m \in M \right\}. \tag{7.10}$$

Then there is a unique vector in N of minimum norm and that this vector is orthogonal to every vector in M.

Proof. According to the above-proved regular version of the finite-dimensional pre-Hilbert space projection theorem, there is a unique vector in M that is closest to $(-x)$: there exists a unique $m_0 \in M$ such that $||m_0 + x|| \leq ||m + x||$ for all $m \in M$. Define $n_0 = (m_0 + x) \in N$; then n_0 is the unique vector in N for which $||n_0|| \leq ||n||$ for all $n \in N$ because, as per (7.10), every vector in N can be written as $(m + x)$ for some $m \in M$. Furthermore, recall that in the proof of the regular version of the theorem it was proved that $(x - m_0)$ is orthogonal to every vector in M. Here, we're using $(-x)$ instead of x so it holds that $(-x - m_0)$ is orthogonal to every vector in M. Switching the sign preserves orthogonality, so $n_0 = (x + m_0)$, the unique vector in N of minimum norm, is orthogonal to every vector in M and the modified version of the theorem is proved. □

7.5 Dual Approximation Theorem

Theorem 5. *Given a real pre-Hilbert space H of finite dimension, a set of k linearly independent vectors $\{y_1, ..., y_k\} \subset H$, and a set of k real numbers $\{\alpha_1, ..., \alpha_k\} \subset \mathbb{R}$, form the set K of vectors $x \in H$ that satisfy the k constraints $(y_1 \mid x) = \alpha_1, ..., (y_k \mid x) = \alpha_k$:*

$$K = \left\{ x \in H \;\middle|\; (y_i \mid x) = \alpha_i,\ i = 1, 2, \dots, k \right\}. \tag{7.11}$$

Then

$$\arg\min_{x \in K} ||x|| = \sum_{i=1}^{k} \beta_i \, y_i \tag{7.12}$$

where

$$\begin{bmatrix} (y_1 \mid y_1) & (y_1 \mid y_2) & \dots & (y_1 \mid y_k) \\ (y_1 \mid y_2) & (y_2 \mid y_2) & \dots & (y_2 \mid y_k) \\ \vdots & \vdots & & \vdots \\ (y_1 \mid y_k) & (y_2 \mid y_k) & \dots & (y_k \mid y_k) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix}. \tag{7.13}$$

In other words, the vector of minimum norm subject to a set of inner product constraints must be a linear combination of the constraint vectors. Furthermore, this solution must exist and is unique: the matrix on the left side of (7.13) is necessarily invertible.

Proof. Let H be n-dimensional and let M be the k-dimensional subspace that has $\{y_1, ..., y_k\}$ as a basis. The orthogonal subspace to M, denoted $M^\perp$, is the $(n-k)$-dimensional subspace of all vectors in H that are orthogonal to M:

$$M^\perp = \left\{ p \in H \;\middle|\; (m \mid p) = 0 \text{ for all } m \in M \right\} = \left\{ p \in H \;\middle|\; (y_i \mid p) = 0 \text{ for all } i \in \{1, ..., k\} \right\}. \tag{7.14}$$

Comparing (7.11) to the right side of (7.14), it's clear that if $\alpha_1 = \alpha_2 = ... = \alpha_k = 0$ then $K = M^\perp$. With arbitrary $\{\alpha_1, ..., \alpha_k\}$, it holds that K is a translation of the subspace $M^\perp$: there exists $v \in H$ such that $K = M^\perp + \{v\} = \{n + v \mid n \in M^\perp\}$. As it turns out, any $v \in K$ does the job. This can be proved by showing that, for all fixed $v \in K$, it holds that $K \subseteq (M^\perp + \{v\})$ and $(M^\perp + \{v\}) \subseteq K$ and thus it must be the case that $K = (M^\perp + \{v\})$.

For all $v, x \in K$, define $p = x - v$. Then, for all $i \in \{1, ..., k\}$, it holds that $(y_i \mid p) = (y_i \mid x - v) = (y_i \mid x) - (y_i \mid v) = \alpha_i - \alpha_i = 0$ where the second to last equality is via (7.11), and thus $p \in M^\perp$ via (7.14). So, $x = p + v \in (M^\perp + \{v\})$. Hence, $K \subseteq (M^\perp + \{v\})$.

For all $v \in K$, $x \in (M^\perp + \{v\})$, it holds by definition of $(M^\perp + \{v\})$ that there exists $p \in M^\perp$ such that $x = p + v$. Then, for all $i \in \{1, ..., k\}$, it holds that $(y_i \mid x) = (y_i \mid p + v) = (y_i \mid p) + (y_i \mid v) = 0 + \alpha_i = \alpha_i$ where the second to last equality is via (7.14) and (7.11), and thus $x \in K$ via (7.11). Hence, $(M^\perp + \{v\}) \subseteq K$.

So, K is the linear variety $K = M^\perp + \{v\}$ where v is any vector in K and M is the subspace that has $\{y_1, ..., y_k\}$ as a basis. We can now apply the modified finite-dimensional pre-Hilbert space projection theorem to K: there exists a unique vector in K of minimum norm and this vector is orthogonal to every vector in $M^\perp$. By definition, the space of vectors orthogonal to $M^\perp$ is M. Thus, the unique vector in K of minimum norm, call it x_0, must be in M and so must be a linear combination of $\{y_1, ..., y_k\}$, which proves (7.12). Furthermore, $x_0 \in K$ so x_0 must satisfy the constraints in (7.11). Substituting (7.12) into these constraints gives

$$(y_i \mid x) = \left(y_i \;\middle|\; \sum_{j=1}^{k} \beta_j y_j \right) = \sum_{j=1}^{k} (y_i \mid y_j)\beta_j = \alpha_i \tag{7.15}$$

for all $i \in \{1, ..., k\}$, which is equivalent to (7.13). The dual approximation theorem is now proved. □

7.6 Geometric Intuition of the Dual Approximation Theorem

Using the notation established in the proof of the dual approximation theorem, let $H = \mathbb{R}^2$ where the inner product is the well-known dot product and orthogonal vectors are at right angles to each

other. As per the proof, $\forall$ fixed $v \in K$, it holds that $K = M^{\perp} + \{v\}$. To establish geometric intuition, specifically choose $v \in M \cap K$ such that $M^{\perp}$, $K = M^{\perp} + \{v\}$ and v are as shown in Figure 1, below. In this example, M is one-dimensional and thus M must be the subspace that has $\{v\}$ as a basis. Accordingly, there can only be one constraint that defines K and it must be of the form $(v \mid x) = \alpha$. We seek the x that satisfies this constraint (i.e. is in K) of minimum norm. From the diagram, it's clear that the vector in K that is closest to the origin is v. So, in this example, $x_0 = v$ is, indeed, a linear combination of the constraint vectors $\{v\}$.

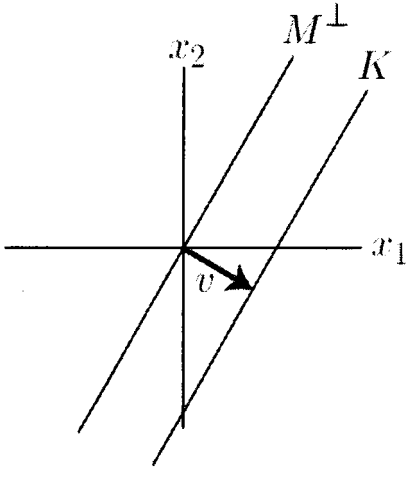


Figure 1: illustration of the dual approximation theorem with one constraint embedded in two dimensions.

This geometric intuition generalizes to arbitrary dimensions. $K = M^{\perp} + \{v\}$ where we choose, without loss of generality, $v \in M \cap K$. M is defined as the subspace that has the constraint vectors as a basis and thus v must be a linear combination of the constraint vectors. Furthermore, it's clear geometrically that, always, $x_0 = v$, and thus the vector in K of minimum norm is a linear combination of the constraint vectors.

7.7 Portfolio Space

Suppose that there are n assets available to the inflation-hedging investor. Let X_i be the random variable that is the rate of return of the i-th asset between today and one year from now. So, if it turns out that $X_i = 0.07$, then \$1 invested today in the i-th asset would be worth \$1.07 in one year. Of course, it is impossible to know exactly what X_i will be—it is a random variable, not a known constant. However, it is assumed that the past several years of asset returns are indicative of asset returns over the upcoming year, so $\mathrm{E}(X_i)$, $\mathrm{var}(X_i)$, and $\mathrm{cov}(X_i, X_j)$, as calculated from historical data, are known for each asset and each pair of assets, respectively.

Represent the investor's portfolio as the n-dimensional vector

$$x = [x_1 \; x_2 \; \ldots \; x_n]^T \tag{7.16}$$

where x_i is the dollar amount of the i-th asset in the portfolio.[4] So, the random variable X that is the rate of return of the entire portfolio can be written

$$X = \sum_{i=1}^{n} x_i X_i \,. \tag{7.17}$$

Let S be the space of all n-dimensional portfolio column vectors; so we have $x \in S$. Let C be the n-by-n covariance matrix of asset returns: the element in the i-th row and j-th column of C is

$$C_{i,j} = \text{cov}(X_i, X_j) \,. \tag{7.18}$$

Via (7.17) and the formula for the variance of a sum of random variables:

$$\text{var}(X) = \text{var}\left(\sum_{i=1}^{n} x_i X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j \,\text{cov}(X_i, X_j) = x^T C x \,. \tag{7.19}$$

Theorem 6. *The following function is a valid inner product over S:*

$$(y \mid z) = y^T C z \quad y, z \in S \,. \tag{7.20}$$

Proof.

- $\text{cov}(X_i, X_j) = \text{cov}(X_j, X_i)$ and thus, via (7.18), C is symmetric, so $[(z \mid y)]^T = (z^T C y)^T = y^T C z = (y \mid z)$. Any 1 by 1 matrix (a scalar) is symmetric so $[(z \mid y)]^T = (y \mid z)$ implies $(y \mid z) = (z \mid y)$ for all $y, z \in S$.

- $(x + y \mid z) = (x + y)^T C z = (x^T + y^T) C z = x^T C z + y^T C z = (x \mid z) + (y \mid z)$ for all $x, y, z \in S$.

- $(\alpha x \mid y) = (\alpha x)^T C y = \alpha x^T C y = \alpha (x \mid y)$ for all $\alpha \in \mathbb{R}$, $x, y \in S$.

- $(x \mid x) = x^T C x = \text{var}(X)$ via (7.19) and variances are nonnegative so $(x \mid x) \geq 0$ for all $x \in S$. Furthermore, $\text{var}(X) = 0$ if and only if X is a constant. It is assumed that there is no nontrivial linear combination of the $\{X_i\}$ that is a constant[5] and thus $\text{var}(X) = 0$ if and only if $x_1 = x_2 = \ldots = x_n = 0$, which is equivalent to x being the zero vector. Thus, $(x \mid x) = 0$ if and only if x is the zero vector.

Thus S is a pre-Hilbert space with the norm

$$||x|| = \sqrt{(x \mid x)} = \sqrt{x^T C x} = \sqrt{\text{var}(X)} \,. \tag{7.21}$$

□

[4]Note that each x_i can be negative (as well as positive, of course) because the investor can short any asset. Shorting an asset is the exact opposite of buying it. If you buy an asset: if its price increases by \$1 then you make \$1 and if its price decreases by \$1 then you lose \$1. If you short an asset: if its price increases by \$1 then you lose \$1 and if its price decreases by \$1 then you make \$1. The mechanics of shorting an asset are as follows: the investor borrows the asset from a lender, immediately sells it to the market, and then buys it back from the market at a later time and returns in to the lender.

[5]If none of the n assets is a financial derivative of one of the others then this "asset independence" is a very safe assumption: there are, in general, no perfect linear relationships between returns of distinct assets so it is impossible to linearly combine distinct assets to create a risk-free return. Furthermore, no asset is risk-free (none of the X_i is a constant) so a single-asset portfolio can never be risk-free.

7.8 Constraints

There are two constraints on x. First, for each \$1 of cash the investor wants to invest, he wants to spend exactly that \$1:

$$\sum_{i=1}^{n} x_i = 1\,. \tag{7.22}$$

Second, the investor wants a return of exactly r on average:

$$\mathrm{E}(X) = \mathrm{E}\left(\sum_{i=1}^{n} x_i\, X_i\right) = \sum_{i=1}^{n} \mathrm{E}(X_i)\, x_i = r \tag{7.23}$$

where the first equality is via (7.17) and the second equality is via the linearity of expectation. In order to facilitate an imminent application of the dual approximation theorem, these constraints are rewritten as inner product constraints in pre-Hilbert space S.

We want to express (7.22) as an equation $(y \,|\, x) = 1$ for some $y \in S$. Such a y would satisfy

$$(y \,|\, x) = y^T C x = \sum_{i=1}^{n} x_i \tag{7.24}$$

Examining the second equality in (7.24), it's clear that we require $y^T C = [1\ 1\ \ldots\ 1]$ and thus $(y^T C)^T = C^T y = Cy = [1\ 1\ \ldots\ 1]^T$ which gives $y = C^{-1}[1\ 1\ \ldots\ 1]^T$. So, the first constraint can be written in inner product form as

$$(y \,|\, x) = \alpha_1\,,\ \ y = C^{-1}[1\ 1\ \ldots\ 1]^T\,,\ \ \alpha_1 = 1\,. \tag{7.25}$$

Similarly, we wish to express (7.23) as an equation $(z \,|\, x) = r$ for some $z \in S$. Such a z would satisfy

$$(z \,|\, x) = z^T C x = \sum_{i=1}^{n} \mathrm{E}(X_i)\, x_i \tag{7.26}$$

Following the same steps as above, we require $z^T C = [E(X_1)\ E(X_2)\ \ldots\ E(X_n)]$ and thus $(z^T C)^T = C^T z = Cz = [E(X_1)\ E(X_2)\ \ldots\ E(X_n)]^T$ which gives

$$z = C^{-1}[E(X_1)\ E(X_2)\ \ldots\ E(X_n)]^T.$$

So, the second constraint can be written in inner product form as

$$(z \,|\, x) = \alpha_2\,,\ \ z = C^{-1}[E(X_1)\ E(X_2)\ \ldots\ E(X_n)]^T\,,\ \ \alpha_2 = r\,. \tag{7.27}$$

Because S is pre-Hilbert there does not exist a nonzero vector $x \in S$ such that $Cx = 0$ (else $(x \,|\, x) = x^T C x = 0$ with $x \neq 0$) and thus C^{-1} necessarily exists.

Note that it's technically possible for y and z to be multiples of each other, which would make the constraints (7.25) and (7.27) either redundant or, more likely, contradictory. However, this only occurs if $[E(X_1)\ E(X_2)\ \ldots\ E(X_n)]$ and $[1\ 1\ \ldots\ 1]$ are multiplies of each other and so, assuming that different assets have different average returns, we ignore the possibility.

7.9 Minimum Variance Portfolio

The investor seeks the portfolio $x \in S$ that minimizes $\text{var}(X) = ||x||^2$ while satisfying the constraints (7.25) and (7.27). The square root function is monotonically increasing so minimizing $||x||$ is equivalent. Thus, the investor seeks the $x \in S$ of minimum norm subject to two inner product constraints.

This is a straightforward application of the dual approximation theorem. Let x_0 be the minimum-variance inflation-hedging portfolio that the investor seeks. Via (7.12), a simple rearrangement of (7.13), and the definition of the portfolio space inner product (7.20), it holds that

$$x_0 = \beta_1 y + \beta_2 z\,, \quad \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} y^T C y & y^T C z \\ y^T C z & z^T C z \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ r \end{bmatrix} \tag{7.28}$$

where C is defined by (7.18) and constructed from historical data for the n assets, y and z are defined in (7.25) and (7.27) respectively, and r is the rate of return demanded by the investor (perhaps the rate of inflation expected to prevail over the next year). Then, via (7.21), the minimum variance is

$$\min \text{var}(X) = ||x_0||^2 = {x_0}^T C x_0\,. \tag{7.29}$$

The fact that this minimum variance portfolio always exists and is unique is not new: the idea is known as the efficient frontier in modern portfolio theory.[6] What is new is the pre-Hilbert space proof of existence and uniqueness presented here.

7.10 Lagrange Multipliers

This solution could have been determined solely with Lagrange multipliers without any mention of norms, inner products, pre-Hilbert space, etc. Accordingly, the pre-Hilbert space method is necessary not because it provides a method for finding the solution if one exists, but rather because it proves that one always exists. Lagrange multipliers are not up to this task because they provide a necessary but not sufficient condition for an extremum.

7.11 Assets Data

Data was gathered for $n = 50$ different assets grouped into three asset classes as listed below.

- Commodities: U.S. consumer price index, Canada consumer price index, Germany consumer price index, U.K. consumer price index, aluminum, live cattle, copper, corn, cotton, gold, live hog, oil, platinum, rubber, silver, soybeans, sugar, wheat.

- Equities: S&P 500 (U.S.), Dow Jones Industrial Average (U.S.), S&P 300 (Canada), CDAX (Germany), TOPIX (Japan), FTSE (U.K.), total NYSE volume (U.S.), and the following U.S. S&P sector indexes: aerospace & defense, auto manufacturing, chemicals, computer hardware, consumable fuels, consumer discretionary, consumer staples, industrials, machinery, packaged foods, pharmaceuticals, retailing, utilities.

- Fixed Income: U.S. 1-year government bonds, U.S. 5-year government bonds, U.S. 10-year government bonds, U.S. 30-year government bonds, Canadian government bonds index, French 10-year government bonds, German 10-year government bonds, Japanese 10-year government bonds, U.K. government bonds index, U.S. corporate bonds index, U.S. AAA-rated corporate bonds index, U.S. municipal bonds index.

[6] http://en.wikipedia.org/wiki/Modern_portfolio_theory#The_efficient_frontier_with_no_risk-free_asset

For each of these 50 assets, annual price data was collected[7] for the 72 years starting with 1940 and ending with 2011, yielding 71 annual return data points for each asset. The asset classes over these 72 years have the following properties. See Appendix 1 for the average return and variance of each asset individually.

asset class	number of assets	overall average return	overall variance
commodities	18	0.069	0.082
equities	20	0.0105	0.055
fixed income	12	0.059	0.008

Table 1: summary of asset classes using annual data from 1940 through 2011.

7.12 Minimum-Variance Inflation Hedge for 2012

For $r = 0.02$ (the expected rate of inflation for 2012), using all 72 years of historical data to calculate $\mathrm{E}(X_i)$ and $\mathrm{cov}(X_i, X_j)$, the full solution x_0 (the minimum-variance inflation hedge) is displayed in Appendix 1. A summary of the solution is as follows:

$$\min \mathrm{var}(X) = 0.00024\,,\quad x_{\text{commodities}} = 0.64\,,\quad x_{\text{equities}} = -0.011\,,\quad x_{\text{fixed income}} = 0.37$$

where $x_{\text{commodities}}$ is the overall fractional investment in all commodity assets, x_{equities} is the overall fractional investment in all equity assets, and $x_{\text{fixed income}}$ is the overall fractional investment in all fixed income assets.

7.13 Benefits of Portfolio Diversification

Examining the $\mathrm{var}(X_i)$ values in Appendix 1, we see that the minimum variance of a 2012 inflation hedge of 0.00024 is smaller than the smallest individual asset variance (which is 0.0011 and belongs to the Germany consumer price index). This is precisely the benefit of portfolio diversification and why this entire exercise is worthwhile. To demonstrate more generally that more available assets does not increase the minimum possible variance,[8] Figure 2, below, shows $\log(\min \mathrm{var}(X))$ for 2012 as a function of n for $r = 0.02$ using all 72 years of historical data. A portfolio of all the assets that data was collected for has the maximum n of 50. For each value of n less than 50, the minimum variance is calculated by performing 100 iterations of randomly removing $(50-n)$ assets and calculating the resultant $\min \mathrm{var}(X)$, and then averaging the 100 values of $\min \mathrm{var}(X)$ that result. Of course, for any given $n \in \{2, 3, ..., 48\}$, this strategy doesn't account for all $\binom{50}{n} > 100$ different portfolios, but, given that $\binom{50}{25} = 1.3 \times 10^{14}$, an approximation must be made for the sake of computational feasibility.

Examining Figure 2, we see that more assets result in a smaller minimum variance (as expected), but with decreasing marginal benefit because the curve is concave up (i.e. going from 9 to 10 assets decreases variance by much more than going from 49 to 50 assets). Note that this effect is lessened by the log scale: $\min \mathrm{var}(X)$ would appear "more" concave in n than $\log(\min \mathrm{var}(X))$ does.

[7]Using the Global Financial Data database (http://www.library.hbs.edu/go/gfd.html).

[8]It would be wrong to say that more available assets necessarily decreases the minimum possible variance because it's possible for an asset to be useless, like one that yields a negative return with unit probability.

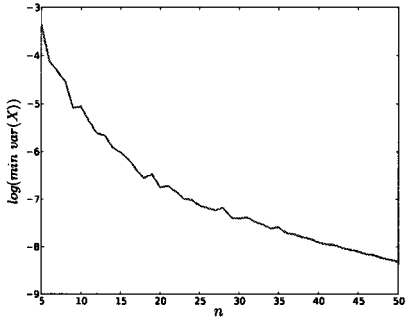


Figure 2: log of the variance of the minimum-variance inflation-hedging portfolio as a function of n, the number of assets available to the investor, with rate of inflation $r = 0.02$.

7.14 Greater Risk, Greater Return

It should also be the case that increasing r, the return required by the investor, should increase the minimum variance. After all, greater returns should necessitate greater risk. Specifically, we examine how the standard deviation of the portfolio, which is $\sqrt{\min \text{var}(X)}$, changes with r. This is a more logical comparison than of $\min \text{var}(X)$ to r because, though $\min \text{var}(X)$, $\sqrt{\min \text{var}(X)}$, and r are all unitless, if r did have units then $\sqrt{\min \text{var}(X)}$ would have the same units whereas $\min \text{var}(X)$ would not. Figure 3, below, shows $\sqrt{\min \text{var}(X)}$ for 2012 as a function of r using all 50 assets and all 72 years of historical data. Also, simply for curiosity's sake, the plot shows how the fractional investments in each asset class ($x_{\text{commodities}}$, x_{equities}, $x_{\text{fixed income}}$) change as the investor demands greater returns r.

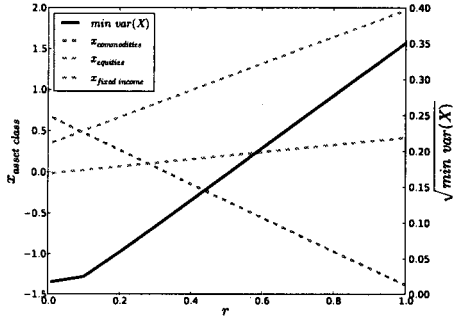


Figure 3: standard deviation of the minimum-variance portfolio that returns r as well as overall fractional investments in the three asset classes all as functions of r (the rate of return demanded by the investor).

As expected, the minimum standard deviation increases with r. Interestingly, the curve appears linear: the investor gets the same return for his standard deviation at any return. This could be interpreted as meaning that all levels of expected return are equally efficient risk-wise.

Also interestingly, the overall fractional investments in the three different asset classes all also appear linear in r. For small r, the optimal portfolio is essentially all commodities and fixed income in roughly equal parts. As r grows large, the investor shorts a large amount of commodities, buys a large amount of fixed income, and buys a comparatively negligible amount of equities. Unsurprisingly, as r grows large the optimal portfolio becomes more leveraged: for each \$1 the investor

wants large return r on, he must short more than \$1 of commodities, leveraging his original \$1 of purchasing power into more than \$2.

Note that the curves in Figure 3 were generated from 11 roughly evenly spaced r values, so the apparent linearity is not simply a result of insufficient r-axis resolution.

7.15 Backtesting

Nothing so far has addressed how well this minimum-variance investment strategy performs in practice. Will the minimum-variance inflation hedge for 2012 actually return roughly $r = 0.02$ over the year? Of course, we can't possibly check this for 2012 until the end of 2012. A method called backtesting is the best alternative. To perform, for example, a backtest for the years 2002 through 2011, we calculate the minimum-variance portfolio for 2002 using only data up to and including 2001 and then calculate how this portfolio would have actually performed in 2002, then we calculate the minimum-variance portfolio for 2003 using only data up to and including 2002 and calculate how this portfolio would have actually performed in 2003, etc. So, a backtest reveals how a strategy that only uses information through the end of year k would have actually performed in year $k + 1$. The results of a backtest for 2002 through 2011 for several different values of r are shown in Table 2 below. The equivalent annual return $a_{\text{equivalent}}$ of each backtest is the constant annual return that would be equivalent to the actual portfolio returns if compounded annually. For example, suppose a backtest yields actual annual returns of a_{2002}, a_{2003}, ..., a_{2011}. For each year there is a different minimum-variance portfolio because another data point is available for each asset. Thus, it is not unreasonable for the investor to, at the end of each year, pull his money out of the old minimum-variance portfolio and put it in the new one, mimicking annual compounding. If he does this, then his cumulative actual return $a_{\text{cumulative}}$ from the beginning of 2002 to the end of 2011 is

$$a_{\text{cumulative}} = \prod_{i=2002}^{2011} (1 + a_i) - 1 . \tag{7.30}$$

This is equivalent to earning an annual return of $a_{\text{equivalent}}$ with annual compounding where

$$(1 + a_{\text{equivalent}})^{10} = 1 + a_{\text{cumulative}} , \tag{7.31}$$

and thus

$$a_{\text{equivalent}} = (1 + a_{\text{cumulative}})^{1/10} - 1 = \left[\prod_{i=2002}^{2011} (1 + a_i) \right]^{1/10} - 1 . \tag{7.32}$$

Quite simply, $(1 + a_{\text{equivalent}})$ is the geometric mean of the $(1 + a_i)$. It's clear from the derivation above that this definition of $a_{\text{equivalent}}$ is the appropriate measure of average annual return.

	$r = 0.02$	$r = 0.04$	$r = 0.06$	$r = 0.08$	$r = 0.1$	$r = 0.2$	$r = 0.3$
a_{2002}	-0.013	0.027	0.066	0.105	0.144	0.34	0.536
a_{2003}	0.052	0.046	0.04	0.033	0.027	-0.005	-0.037
a_{2004}	0.042	0.052	0.061	0.071	0.08	0.129	0.177
a_{2005}	0.257	0.233	0.209	0.185	0.16	0.039	-0.081
a_{2006}	0.014	0.04	0.066	0.093	0.119	0.25	0.382
a_{2007}	-0.006	0.018	0.043	0.068	0.093	0.218	0.342
a_{2008}	0.058	0.012	-0.034	-0.08	-0.127	-0.358	-0.59
a_{2009}	-0.006	0.055	0.115	0.176	0.236	0.538	0.84
a_{2010}	-0.102	-0.063	-0.024	0.015	0.055	0.251	0.447
a_{2011}	0.042	0.026	0.009	-0.007	-0.023	-0.103	-0.184
$a_{\text{equivalent}}$	0.03	0.042	0.053	0.063	0.072	0.102	0.102

Table 2: annual returns for 2002 through 2011 backtests with various values of r.

These results show that, for a wide range of r values, the minimum-variance portfolio is not reliable over a single year. For example, though $r = 0.02$ yields $a_{\text{equivalent}} = 0.03 \approx 0.02$, the individual returns for $r = 0.02$ over the ten years of backtesting range from -0.1 to .26 with no consistency around 0.02. In other words, it seems unlikely that the minimum-variance inflation hedge for 2012 will actually be an inflation hedge! It could very well return well above or well below the desired level of $r = 0.02$. Thus, the method presented in this paper is not useful with regards to the initial goal of a one-year inflation hedge.

However, we note from Table 2 that, at least for the smaller values of r, it holds that $a_{\text{equivalent}} \approx r$. So, while the minimum-variance portfolio may not be effective over a single year, investing in the minimum-variance portfolio each year for 10 years yields roughly the demanded annual return. This observation lead to the author pursuing a senior thesis in which he developed a practical and profitable investment strategy based on the minimum-variance portfolio.

7.16 Acknowledgments

My inspiration for solving the minimum variance problem using the pre-Hilbert space method presented here was a homework problem for Math 116 written by Paul Bamberg, the course's instructor. Math 116 covers the dual approximation theorem and Paul had the idea of applying it to construct minimum variance portfolios given fixed portfolio expectation and known individual asset expectations and covariances.

7.17 References

Luenberger, David. Optimization by Vector Space Methods. New York: Wiley, 1968.

http://en.wikipedia.org/wiki/Modern_portfolio_theory#The_efficient_frontier_with_no_risk-free_asset

Appendix 1: Full Solution x_0 for $r = 0.02$ (For 2012 using all 72 years of historical data.)

	fractional investment x_i	asset	average return $\mathrm{E}(X_i)$	return variance $\mathrm{var}(X_i)$	
commodities	-0.0534	aluminum	0.0431	0.0415	
	0.3779	Canada consumer price index	0.0391	0.0012	← 3rd largest buy
	0.0289	cattle	0.0439	0.0169	
	0.0581	copper	0.0841	0.0954	
	-0.0232	corn	0.0638	0.0738	
	-0.0085	cotton	0.0566	0.0756	
	0.5275	Germany consumer price index	0.0288	0.0011	← 2nd largest buy
	-0.032	gold	0.0775	0.0568	
	0.0202	hog	0.0689	0.0908	
	-0.0164	oil	0.0972	0.1123	
	-0.0444	platinum	0.0749	0.0454	
	0.0178	rubber	0.102	0.2516	
	0.0137	silver	0.1191	0.229	
	0.0274	soybeans	0.0613	0.0599	
	0.0087	sugar	0.1324	0.2455	
	-0.2162	U.K. consumer price index	0.0512	0.0022	← 3rd largest short
	-0.0486	U.S. consumer price index	0.0404	0.0012	
	0.0029	wheat	0.0586	0.0774	
equities	-0.0659	S&P300 (Canada)	0.116	0.0263	
	0.0935	Dow Jones Industrial Average (U.S.)	0.0777	0.0245	
	0.0233	CDAX (Germany)	0.1222	0.1009	
	-0.0108	TOPIX (Japan)	0.1551	0.1054	
	0.0342	S&P aerospace & defense	0.1121	0.0819	
	-0.0208	S&P auto manufacturing	0.1013	0.1688	
	-0.0094	S&P chemicals	0.0785	0.0338	
	0.0068	S&P computer hardware	0.1375	0.0735	
	-0.0369	S&P consumable fuels	0.1052	0.0293	
	0.1117	S&P consumerdiscretionary	0.0916	0.0365	
	-0.0039	S&P consumerstaples	0.0933	0.0295	
	-0.1647	S&P industrials	0.0827	0.0313	
	0.0213	S&P machinery	0.0996	0.0515	
	0.0174	S&P packagedfoods	0.0952	0.03	
	-0.0574	S&P pharmaceuticals	0.1071	0.0412	
	-0.0902	S&P retailing	0.1084	0.0572	
	0.0335	S&P utilities	0.0589	0.0321	
	0.1538	S&P500	0.0831	0.0274	
	-0.0365	FTSE (U.K.)	0.1456	0.0639	
	-0.0102	NYSE volume (U.S.)	0.1282	0.061	
fixed income	-0.4128	Canadian government bonds index	0.0506	0.0017	← 2nd largest short
	0.0659	French 10-year government bonds	0.0695	0.008	
	0.0926	German 10-year government bonds	0.0551	0.0179	
	-0.0563	Japanese 10-year government bonds	0.064	0.016	
	0.0872	U.K. government bonds index	0.0605	0.0016	
	-0.0403	U.S. AAA-rated corporate bonds index	0.0684	0.0056	
	-0.1494	U.S. corporate bonds index	0.0743	0.0069	
	0.2326	U.S. municipal bond index	0.0421	0.0054	
	-0.0267	U.S. 10-year government bonds	0.0602	0.0076	
	0.9741	U.S. 1-year government bonds	0.0475	0.0014	← largest buy
	0.0477	U.S. 30-year government bonds	0.0627	0.0166	
	-0.4435	U.S. 5-year government bonds	0.0548	0.0041	← largest short

FEATURE

8

MY FAVORITE PROBLEM

Poncelet's Porism

François Greer[†]
Harvard University '11
Cambridge, MA 02138
fgreer@post.harvard.edu

8.1 The Porism

Our story begins with a quaint problem in plane geometry.

Let C and D be ellipses in $\mathbb{R}^2$, D inside C. Show that if there exists an n-gon simultaneously circumscribed on D and inscribed in C (Fig. 1), then there exists a 1-parameter family of n-gons with this property, whose vertices sweep out C. Remark: if we allow the n-gons to be self-intersecting (Fig. 2), then in fact the turning number is constant as well.

Fig. 1

Fig. 2

More explicitly, choose a point $p \in C$. There are two lines through p tangent to D. Let ℓ be the right hand line (from the perspective of p). Set $T(p) \in C$ be the other point where ℓ meets C. Iterating this transformation of C, we obtain a sequence of points $\{T^n(p)\}_{n\in\mathbb{N}}$. We must show that the periodicity of this sequence depends only on C and D, not on the starting point, p. Slightly awkward to state, but geometrically quite compelling, this classical fact is known as **Poncelet's Porism**[1]. It was first proved in 1814 by French engineer and mathematician Jean-Victor Poncelet, who was a prisoner-of-war in Russia at the time, following Napoleon Bonaparte's failed invasion. We present two proofs: the first uses elementary measure theory, and the second is a brilliant application of the classical theory of algebraic curves. For reasons of accessibility, we adopt an informal tone for the more technical aspects; the interested reader is encouraged to fill in the details.

8.2 Following Your Nose

Let us first consider the case where C and D are circles centered at the origin with radii 1 and r, respectively. If $\theta(p) \in \mathbb{R}/2\pi\mathbb{Z}$ denotes the angle from the x-axis to p and $T(p) = q$, then basic trigonometry tells us that $\Delta\theta := \theta(q) - \theta(p) = \pi - 2\arcsin(r)$.

[†]François Greer is a graduate student at Stanford studying mathematics.

[1]A lesser known synonym of the word *theorem*, *porism* was featured in the title of a lost treatise by Euclid. Poncelet's use of the word was surely motivated by the Euclidean geometric flavor of the problem.

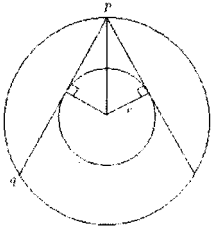


Fig. 3

Since the sequence $\{\theta(T^n(p))\}_{n\in\mathbb{N}}$ is arithmetic, the period of $\{T^n(p)\}_{n\in\mathbb{N}}$ is the smallest integer k such that $k\Delta\theta \in 2\pi\mathbb{Z}$. The value of $\Delta\theta$ is independent of p, so the porism follows. Emboldened by this example, we attempt the general case. After translating, rotating, and scaling the axes, we may assume that C is the unit circle, since these invertible operations preserve ellipses and lines. For general D, the value of $\Delta\theta$ is no longer constant, so instead we seek a measure $d\tilde{\theta} = g(\theta)\cdot d\theta$ on C that *is* T-invariant, or equivalently a positive, integrable function $g\colon C \to \mathbb{R}$ such that

$$\int_a^b g(\theta)\,d\theta = \int_{T(a)}^{T(b)} g(\theta)\,d\theta$$

In particular, we obtain that

$$\Delta\tilde{\theta} := \int_p^{T(p)} g(\theta)\,d\theta$$

is independent of p. Since any interval $[a,b] \subset \mathbb{R}/2\pi\mathbb{Z}$ can be decomposed into many disjoint smaller intervals, it suffices to check the T-invariance condition for arbitrarily small intervals. Since g is continuous, as $\Delta\theta \to 0$, we want to show the infinitessimal relation

$$g(p)\cdot\Delta\theta = g(q)\cdot\Delta\psi$$

Fig. 4

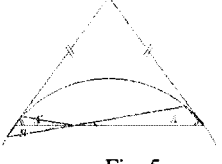


Fig. 5

When $\Delta\theta$ is small, the arc length is linearized, so we may replace arcs with their corresponding chords (Fig. 4). Now, if $|p, D|_R$ denotes the distance from p to the right hand tangent point, and $|q, D|_L$ denotes the distance from q to the left hand tangent point, we want

$$\frac{\Delta\theta}{\Delta\psi} = \frac{|p, D|_R}{|q, D|_L}$$

because triangle A is similar to triangle A', and the left hand sides of triangles A' and B are equal in the limit, by the small angle approximation (Fig. 5). Thus, we want

$$g(p)\cdot|p, D|_R = g(q)\cdot|q, D|_L.$$

The presence of both left and right in the equation above prevents us from defining $g(x)$ as $|x, D|^{-1}$. To resolve this asymmetry, we choose some linear transformation $A\colon \mathbb{R}^2 \to \mathbb{R}^2$ sending D to a circle. Then

$$\frac{|p, D|_R}{|q, D|_L} = \frac{|A(p), A(D)|_R}{|A(q), A(D)|_L} = \frac{|A(p), A(D)|}{|A(q), A(D)|}$$

because the numerator and denominator of the LHS are collinear measurements. We dropped the subscripts in the last equality because the two notions agree, since $A(D)$ is a circle. Now, we can at last define

$$g(p) = \frac{1}{|A(p), A(D)|},$$

which satisfies the desired local invariance.

8.3 Complex Geometry

The second method displays a great deal more foresight, and establishes the result in a more general setting. As such, it serves as a pleasant advertisement for projective algebraic geometry over $\mathbb{C}$. The first step is to consider the defining equation $f(x, y)$ (resp. $g(x, y)$) of C (resp. D) as polynomials over the larger field $\mathbb{C} \supset \mathbb{R}$. The complex solutions form a locus in $\mathbb{C}^2$, which we will again call C (resp. D), abusing notation. Next, we homogenize the equations

$$F(X, Y, Z) := Z^2 f(X/Z, Y/Z) \quad \text{and} \quad G(X, Y, Z) := Z^2 g(X/Z, Y/Z)$$

whose complex solutions form loci in $\mathbb{CP}^2 = (\mathbb{C}^3 - \{0\})/\mathbb{C}^\times$, a natural compactification of $\mathbb{C}^2$. These loci are examples of algebraic curves.

Definition 1. A smooth, projective algebraic curve is the zero locus in $\mathbb{CP}^n$ of a system of homogeneous polynomials in $n + 1$ variables, which is a (complex) submanifold[2] of dimension 1. Henceforth, we call such objects "curves." NB: Dimension 1 over $\mathbb{C}$ means dimension 2 over $\mathbb{R}$.

This is the proper setting for the following classical theorem:

Theorem 2 (Bézout's Theorem). *If X and Y are distinct curves in $\mathbb{CP}^2$ cut out by irreducible polynomials of degrees d and e, then they intersect in de points, counted with multiplicity.*

In our case, C and D have 4 intersection points, none of which lie in $\mathbb{R}^2$. Furthermore, Bèzout allows us to get at the topology of our conics:

Proposition 3. *A smooth conic curve $C \subset \mathbb{CP}^2$ is isomorphic to $\mathbb{CP}^1$, the Riemann sphere.*

Proof. Choose a point $x \in C$, and a line $L \simeq \mathbb{CP}^1$ in $\mathbb{CP}^2$ with $x \notin L$. Define a map $\phi\colon C \to L$ sending a point $y \in C$ to the intersection of the line $\overline{xy}$ with L. This map is bijective by Bézout's Theorem because every line through x meets C in exactly one other point (which may be equal to x, meaning that the line is tangent to C). One can check that both ϕ and its inverse are given by polynomial expressions, so ϕ defines an isomorphism of curves. □

Next, consider the space

$$\Omega = \{(p, \ell) : p \in \ell,\ p \in C,\ \ell \text{ tangent to } D\} \subset \mathbb{CP}^2 \times (\mathbb{CP}^2)^*$$

where $(\mathbb{CP}^2)^*$ denotes the set of lines in $\mathbb{CP}^2$. Poncelet's Porism can be reduced to studying the geometry of Ω, which for now is just a topological space. Indeed, define involutions $\alpha, \beta\colon \Omega \to \Omega$ as

$$\alpha(p, \ell) = (p', \ell),$$

where p' is the other point where ℓ meets C, or just p again if ℓ is tangent to C, and

$$\beta(p, \ell) = (p, \ell'),$$

where ℓ' is the other line through p tangent to D, or just ℓ again if $p \in D$. Note that both α and β have fixed points. The key is that transformation $\tau = \beta \circ \alpha$ acts as T on the first coordinate! Thus, our problem reduces to studying the iterates of τ. Before proceeding, we add more geometric

[2]Every algebraic curve is a compact Riemann surface (1-dimensional complex manifold). The converse is true as well: every Riemann surface admits an embedding into $\mathbb{CP}^n$, where it is cut out by homogeneous polynomials.

structure to the space Ω. We have a continuous map $\pi\colon \Omega \to C$ given by projection onto the first coordinate. Since each point of C not on D lies on two different tangent lines to D, this map is two-to-one away from the 4 points of $C \cap D$, where it is one-to-one. In fact, if we remove these 4 points from C and their pre-images from Ω, we obtain a topological covering map $\pi|_{\Omega'}\colon \Omega' \to C'$. We will use this covering to give Ω the structure of a compact Riemann surface. Recall that any nonconstant holomorphic map of compact Riemann surfaces $p\colon X \to Y$ can be expressed in suitably chosen local coordinates z on X and w on Y as

$$w = z^m.$$

When $m \neq 1$, the point corresponding to $w = 0$ is called a *branch point* of the map, and $m - 1$ the degree of ramification. If we remove the set branch points (which is finite by compactness) from Y and their fibers from X, then p restricts to a topological covering.

Theorem 4 (Riemann Existence). *Let Y be a compact Riemann surface, and X° be a topological space with a finite covering map $p\colon X^\circ \to Y^\circ = Y - \{y_1, \ldots, y_b\}$. Then, for any given monodromy data — i.e. an action of $\pi_1(Y, y_0)$ on the fiber $p^{-1}(y_0)$, for some $y_0 \neq y_i$ for any $i = 1, 2, \ldots, b$ — there is a unique smooth and compact Riemann surface X. obtained by adding finitely many points to X°, and holomorphic map $\tilde{p}\colon X \to Y$, branched over $y_1, \ldots, y_b$ and extending p.*

One can check that our topological covering $\pi : \Omega \to C$ has nontrivial monodromy around the 4 removed points, so Riemann Existence produces a compact, connected Riemann surface homeomorphic to Ω over C.

Theorem 5 (Riemann-Hurwitz Formula). *If $\pi\colon X \to Y$ is a nonconstant map of compact Riemann surfaces of genera g and h respectively, then*

$$(2g - 2) = d(2h - 2) + r$$

where r is the sum of all ramification degrees, and d is the size of an unramified fiber.

In our case, $\pi\colon \Omega \to C$ has 4 ramification points, each with degree 1. Since $C \simeq \mathbb{CP}^1$ has genus 0, we deduce that Ω must have genus 1. We also observe that $\alpha\colon \Omega \to \Omega$ is holomorphic, since it simply interchanges the sheets of the cover π. In a suitable chart around each ramification point, we have $w = z^2$, so $\alpha(z) = -z$.

At this point, we note that our choice of $\pi : \Omega \to C$ was rather asymmetric. If instead we considered the projection of Ω onto the second coordinate, we would obtain a ramified cover $\pi'\colon \Omega \to D^* \subset (\mathbb{P}^2)^*$. A quick check reveals that D^* is a conic in $(\mathbb{P}^2)^*$, so the exact same argument holds in this dual setting. The map β interchanges the sheets of the covering π', so it is also holomorphic. To wrap up, we appeal to one last fact, a special case of the Uniformization Theorem.

Fact 6. *Every genus 1 compact Riemann surface is isomorphic to $\mathbb{C}/\Lambda$, where $\Lambda \subset \mathbb{C}$ is a lattice.*

In particular, $\Omega \simeq \mathbb{C}/\Lambda$ now has a commutative group law, given by addition in $\mathbb{C}/\Lambda$. This allows us to describe the involutions α and β more explicitly.

Proposition 7. *All holomorphic involutions ι of $\mathbb{C}/\Lambda$ are given by $[z] \mapsto [\pm z + t]$, for $t \in \mathbb{C}$.*

Proof. An isomorphism of Riemann surfaces $\iota\colon \mathbb{C}/\Lambda \to \mathbb{C}/\Lambda$ lifts to an isomorphism of universal covers $\mathbb{C} \to \mathbb{C}$, by the universal property of universal covers. By standard complex analysis, any invertible holomorphic function $\mathbb{C} \to \mathbb{C}$ is an affine transformation $z \mapsto az + b$ $(a, b \in \mathbb{C}, a \neq 0)$. The involution condition requires that

$$\iota^2(z) = a(az + b) + b = a^2 z + b(a + 1) = z \in \mathbb{C}/\Lambda$$

$$\Rightarrow (a^2 - 1)z + b(a + 1) \in \Lambda$$

for all $z \in \mathbb{C}$. This is only true if $a = \pm 1$. □

Since $\alpha, \beta \colon \Omega \to \Omega$ have fixed points, and neither is the identity map, $a = -1$. Thus, we can write

$$\begin{aligned} \alpha([z]) &= [-z + t] \\ \beta([z]) &= [-z + s] \\ \Rightarrow \tau([z]) &= [z - t + s] \end{aligned}$$

Therefore, τ acts on Ω via translation by $[s - t] \in \mathbb{C}/\Lambda$. If $[s - t]$ is a torsion point (of order say m) with respect to the group law, then the sequence $\{T^n(p)\}_{n \in \mathbb{N}}$ is m-periodic, regardless of the starting point. If $[s - t]$ has infinite order, then the sequence never repeats, again regardless of the starting point. We have thus established Poncelet's Porism in the more general context of conics in $\mathbb{CP}^2$, where there is no requirement that lie D within C (in fact, being conics in $\mathbb{CP}^2$, they must intersect).

References

[1] Flatto, Leopold. Poncelet's Theorem. Providence: AMS, 2009.

[2] Griffiths, P. & Harris, J. "On Cayley's Explicit Solution to Poncelet's Porism." *L'Enseignement Math.* 24 (31-40), 1978.

[3] King, J. L. "Three Problems in Search of a Measure." *Amer. Math. Monthly.* 101 (609-628), 1994.

FEATURE

9

Problems

The HCMR welcomes submissions of original problems in any fields of mathematics, as well as solutions to previously proposed problems. Proposers should direct problems to `hcmr-problems@hcs.harvard.edu` or to the address on the inside front cover. A complete solution or a detailed sketch of the solution should be included, if known. Unsolved problems will *not* be accepted. Solutions to previous problems should be directed to `hcmr-solutions@hcs.harvard.edu` or to the address on the inside front cover. Solutions should include the problem reference number, the solver's name, contact information, and affiliated institution. Additional information, such as generalizations or relevant bibliographical references, is also welcome. Correct solutions will be acknowledged in future issues, and the most outstanding solutions received will be published. To be considered for publication, solutions to the problems below should be postmarked no later than *December 24, 2012*. We encourage all submitters to typeset their submissions in LaTeX and submit the source code along with the pdf.

A12 – 1. Evaluate

$$\lim_{n\to\infty} \prod_{k=1}^{n} \frac{\cosh(k^2+k+\frac{1}{2}) + i\sinh(k+\frac{1}{2})}{\cosh(k^2+k+\frac{1}{2}) - i\sinh(k+\frac{1}{2})}.$$

Proposed by Moubinool Omarjee (Paris, France)

A12 – 2. Let $f : [0,1] \to [0,\infty)$ be an integrable function which is left continuous at 1. Find the value of

$$\lim_{n\to\infty} n \int_0^1 \left(\sum_{k=n}^{\infty} \frac{x^k}{k} \right) f(x)\, dx.$$

Proposed by Ovidiu Furdui (University of Toledo, Cluj, Romania)

A12 – 3. Let a, b, c be positive real numbers. Prove that

$$\frac{16}{27}\left(\frac{a}{b+c} + \frac{b}{c+a} + \frac{c}{a+b}\right)^3 + \sqrt[3]{\frac{abc}{(a+b)(b+c)(c+a)}} \geq \frac{5}{2}.$$

Proposed by Tuan Le (Anaheim, CA)

A12 – 4. For n a positive integer, evaluate

$$\int_0^1 \cdots \int_0^1 \lfloor x_1 + \cdots + x_n \rfloor \, dx_1 \cdots dx_n.$$

Proposed by Yale Fan (Harvard College '14)

Editor's Note: The following problems from the previous issues are released again as they received no solutions.

A11 – 3. Let $E = \{M \in \mathrm{Mat}_{3\times 3}(\mathbb{R}) : tr(M) = 0 \text{ and } 4(tr(M^*))^3 + 27(\det(M))^2 > 0\}$ where M^* is the adjugate matrix of M. Let $A, B \in E$ such that A and B have no common eigenvectors. *Suppose*

$$\langle Be_1, e_3\rangle\langle Be_2, e_1\rangle\langle Be_3, e_2\rangle = \langle Be_1, e_2\rangle\langle Be_2, e_1\rangle\langle Be_3, e_1\rangle$$

where $\langle , \rangle$ denotes the inner product and (e_1, e_2, e_3) is the canonical basis of $\mathbb{R}^3$. Suppose as well that

$$A^{n_1}B^{q_1}A^{n_2}B^{q_2}\cdots A^{n_k}B^{q_k} = I$$

where $n_i, q_i \in \mathbb{Z}$. Prove that

$$A^{-n_1}B^{-q_1}A^{-n_2}B^{-q_2}\cdots A^{-n_k}B^{-q_k} = I$$

Proposed by Moubinool Omarjee (Paris, France).

A11 – 4. Let x, y, z be three positive real numbers such that $x + y + z = xyz$. Prove that

$$\sum_{\text{cyc}} \frac{1}{\sqrt{x^2+1}} \leq \sum_{\text{cyc}} \frac{1}{x^2+1} + \sum_{\text{cyc}} \frac{1}{\sqrt{(x^2+1)(y^2+1)}} \leq \frac{3}{2}$$

Proposed by Cezar Lupu (University of Bucharest, Bucharest, Romania).

A11 – 5. Let $a \in \mathbb{N}^*$ be a fixed integer. Prove that there are an infinity of positive integers m such that $\sigma(am) < \sigma(am+1)$ where $\sigma(n)$ is the sum of the divisors of the positive integer n.

Proposed by Vlad Matei (University of Bucharest, Bucharest, Romania).

A11 – 6. Consider the set S of all strings over an alphabet of three symbols. Give it a group structure where the law of composition is concatenation and each word w in the group satisfies the relation $ww = 1$. Compute the structure of the group and show that although it is finite there exists an infinite string with no substring of the form ww, where w is a word.

Note. There is an interesting generalization of this problem by replacing the relation $w^2 = 1$ with higher powers. We encourage the interested reader to submit his or her solution to this generalization as well.

Proposed by Lucia Mocz '13 and Dmitry Vaintrob '11.

S08 – 2. Professor Perplex is at it again! This time, he has gathered his $n > 0$ combinatorial electrical engineering students and proposed:

> "I have prepared a collection of $r > 0$ identical *and indistinguishable* rooms, each of which is empty except for $s > 0$ switches *all initially set to the 'off' position.* You will be let into the rooms at random, in such a fashion that no two students occupy the same room at the same time and every student will visit each room arbitrarily many times. Once one of you is inside a room, he or she may toggle any of the s switches before leaving. This process will continue until some student chooses to assert that all the students have visited all the rooms at least $v > 0$ times each. If this student is right, then there will be no final exam this semester. Otherwise, I will assign a week-long final exam on the history of the light switch."

What is the minimal value of s (as a function of n, r, and v) for which the students can guarantee that they will not have to take an exam?

Proposed by Scott D. Kominers '09/AM'10/PhD'11, Paul Kominers (MIT '12), and Justin Chen (Caltech '09).

FEATURE

10 Solutions

Convexity

A11 – 1. Let a, b, c be positive real numbers. Prove that:

$$\frac{\sqrt{a^3+b^3}}{a^2+b^2}+\frac{\sqrt{b^3+c^3}}{b^2+c^2}+\frac{\sqrt{c^3+a^3}}{c^2+a^2}\geq\frac{6(ab+bc+ac)}{(a+b+c)\sqrt{(a+b)(b+c)(c+a)}} \tag{10.1}$$

Proposed by Tuan Le (Fairmont High School, Anaheim, CA)

Solution by Paolo Perfetti. We need the following result:

$$(a+b)(b+c)(c+a)\geq\frac{8}{9}(a+b+c)(ab+bc+ca).$$

This follows from the equality $(a+b)(b+c)(c+a)=(a+b+c)(ab+bc+ca)-abc$, and from the fact that

$$(a+b+c)(ab+bc+ca)\geq 9abc.$$

The last inequality follows by AM-GM since $(a+b+c)(ab+bc+ca)\geq 3(abc)^{1/3}3(abc)^{2/3}=9abc$.

Now by $a^2+b^2\geq 2ab$, we have $\sqrt{a+b}\sqrt{a^3+b^3}\geq a^2+b^2$; thus the inequality (10.1) is implied by

$$\frac{1}{\sqrt{a+b}}+\frac{1}{\sqrt{b+c}}+\frac{1}{\sqrt{c+a}}\geq\frac{6(ab+bc+ca)}{(a+b+c)\sqrt{(a+b)(b+c)(c+a)}}.$$

Now the convexity of $1/\sqrt{x}$ allows us to write

$$\frac{1}{\sqrt{a+b}}+\frac{1}{\sqrt{b+c}}+\frac{1}{\sqrt{c+a}}\geq\frac{3\sqrt{3}}{\sqrt{2}\sqrt{a+b+c}}.$$

Therefore, we only need to prove that:

$$\frac{3\sqrt{3}}{\sqrt{2}\sqrt{a+b+c}}\geq\frac{6(ab+bc+ca)}{(a+b+c)\sqrt{(a+b)(b+c)(c+a)}}$$

By cross-multiplying and squaring, this is equivalent to

$$3(a+b+c)^2(a+b)(b+c)(c+a)\geq 8(ab+bc+ca)^2(a+b+c)\iff$$
$$3(a+b+c)(a+b)(b+c)(c+a)\geq 8(ab+bc+ca)^2.$$

The result above yields

$$3(a+b+c)(a+b)(b+c)(c+a)\geq 3(a+b+c)\frac{8}{9}(a+b+c)(ab+bc+ca)\geq$$
$$8(ab+bc+ca)^2\iff(a+b+c)^2\geq 3(ab+bc+ca).$$

But this follows from the trivial result $a^2+b^2+c^2\geq ab+bc+ca$, which in turn follows from $(a-b)^2+(b-c)^2+(c-a)^2\geq 0$. □

Editor's Note: Paolo Perfetti points out that this problem also appeared as

- http://www.math.ust.hk/excalibur/v14_n3.pdf, problem 3 and
- http://ssmj.tamu.edu/problems/October-2010.pdf, problem 5107.

Those Left Out Form Their Own Group

A11 – 2. Are there any simple groups of order $p(p+1)$, where p is prime?

Proposed by Eric Larson '13.

Solution by Evan O'Dorney and Allen Yuan.

Suppose we have a simple group G of order $p(p+1)$.

The number of Sylow p-subgroups is more than 1, is a divisor of $p+1$, and is 1 mod p. Therefore, there are exactly $p+1$ of them.

The elements of G can be divided into three categories: (1) the identity, (2) $(p+1)(p-1)$ elements of order p, which we will call "ordinaries," (3) the remaining p elements, which we will call "extras." We will aim to prove that the extras, together with the identity, form a group, which must obviously be normal.

Let us find a lower bound on the number of equations of the form $ab = c$, where a and b are ordinary and c is an extra. Fix two distinct Sylow p-subgroups A and B, and let a be a non-identity element of A. Assume that ab is ordinary (it clearly cannot be the identity) for all non-identity b in B. It is not hard to verify that such ab cannot lie in A, B, or aBa^{-1}, and that these three Sylow p-subgroups are different. So the $p-1$ values of ab lie in the $(p+1)-3 = p-2$ remaining Sylow p-subgroups. Thus, we deduce that two of them, namely ab and ab', lie in the same subgroup C. Then $b^{-1}b'$ is in C, which is a contradiction.

Using this method, we can get $(p+1)p(p-1)$ such equations $ab = c$ ($p+1$ choices for A, p for B, $p-1$ for a). By construction, no two are alike, so no two have the same (a, c) pair. The number of possible (a, c) pairs is also $(p+1)p(p-1)$ (a is one of $(p+1)(p-1)$ ordinaries, c is one of p extras). We conclude that the equation $ab = c$ has no solutions where a is ordinary and b and c are extras. Passing to the equation $a = cb^{-1}$, we derive that the product of two extras is never ordinary. So the extras, together with the identity, are closed under multiplication, as desired. □

FEATURE

10

ENDPAPER

Waiting for Mathematics

Professor Gerald E. Sacks [†]
Harvard University
Cambridge, MA 02138
sacks@math.harvard.edu

On several occasions the solution to a mathematical problem came to me while waiting.

The first time occurred while waiting at the end of a long line to the checkout counter in a supermarket near Princeton. For what seemed like the hundredth time, I struggled with the last part of my Ph.D. thesis: is there a minimal Turing degree below zero prime? The answer came just before my turn to pay for some cream cheese. Yes! Simply apply the finite injury method to the construction of Spector trees. If the line to the checkout counter had any been shorter, so would have been my thesis. (To this day, I am truly fond of cream cheese.) This was the first time I did mathematics *on line*.

Two years ago my friend P____ was late for an appointment in front of his apartment building in Manhattan. I idly considered the last remaining obstacle in my proposed proof of density for the recursively enumerable Turing degrees. What was needed was a way to "remove" an element of a recursively enumerable set A added at an earlier stage of the enumeration of A. On the surface that seemed impossible. Fortunately P____ was quite late. As he approached, I smiled. Naturally, I was glad to see him. But my elation also came about from the fact I had realized the effect of a "removal" could be achieved by adding infinitely many elements of a certain type.

Another instance of mathematical progress occurred while sweating out a long and turbulent plane landing. I would not like to do that again. The theorem was not worth it.

Waiting appears to focus the mind in unexpected ways. Perhaps it reduces external disturbances. My experiences with waiting lead me to believe I am doing mathematician X a favor if I am late for an appointment with X.

[†]Professor Gerald E. Sacks is a member of the Harvard Department of Mathematics.