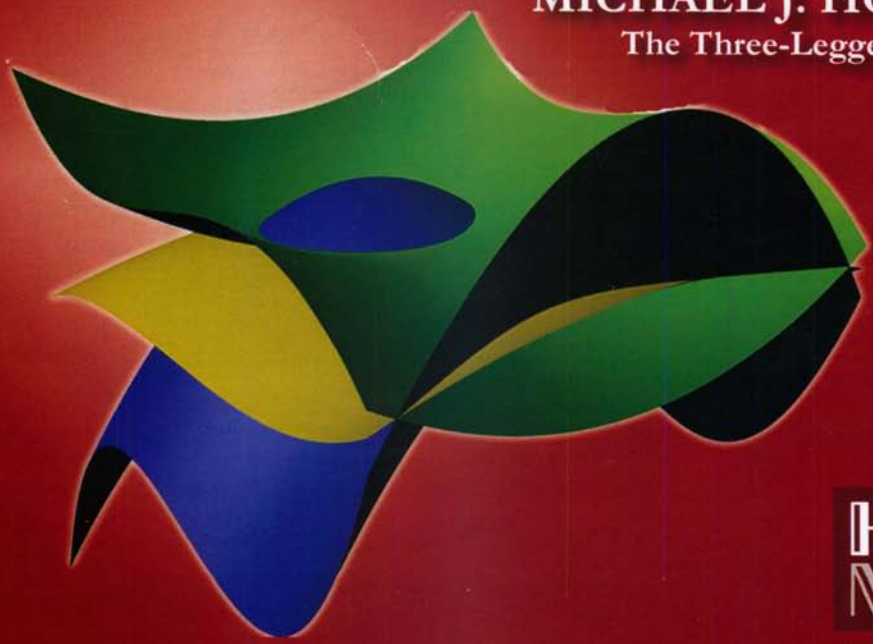
The Harvard College Mathematics Review

Volume 3

Spring 2011

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 Lucia Mocz 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 Gerishom Gimaiyo.

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 · Jane Street Capital · The Harvard University Mathematics Department

Cover Image. The image on the cover depicts several functions, whose common zero locus describes an algebraic variety (which is in this case an elliptic curve). This issue's article "Hilberts Nullstellensatz and Schemes" by Miles Edwards (p. 17) describes one of the foundational theorems in algebraic geometry, relating an ideal in a polynomial ring to its corresponding algebraic variety. This image was created in Asymptote by Graphic Artists Eric Larson and Katherine Banks.

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

-2
Contents

-1
Staff

Editor-In-Chief
Rediet Abebe '13

Business Manager
Gerishom Gimaiyo '13

Articles Editor
Francois Greer '11
Features Editor
Katherine Banks '12
Problems Editor
Lucia Mocz '13

Associate Director
Eric Larson '13
Assistant Articles Editor
Akhil Mathew '14
Assistant Problems Editors
Ge Yang '14 and Levent Alpoge '14

Issue Production Directors
Eric Larson '13 and Katherine Banks '12

Graphic Artists
Eric Larson '13 and Katherine Banks '12

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

Webmaster
Rediet Abebe '13

Board of Reviewers
Levent Alpoge '14
Anirudha Balasubramanian '14
Katherine Banks '12
John Casale '12
Ashok Cutkosky '13
Yale Fan '14
Francois Greer '11
Eric Larson '13
Geoffrey Lee '14
Akhil Mathew '14
Lucia Mocz '13
Ge Yang '14
Zachary Young '14

Board of Copy Editors
Rediet Abebe '13
John Casale '12
Ashok Cutkosky '13
Yale Fan '14
Francois Greer '11
Eric Larson '13
Geffory Lee '14
Akhil Mathew '14
Lucia Mocz '13
Zachary Young '14

Faculty Adviser
Professor Peter Kronheimer, Harvard University

0 From the Editor

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

When I took on the role of Editor-In-Chief of the Harvard College Math Review in May 2010, I was not sure how HCMR was going to turn out, or if we were even going to have an HCMR by the end of the school year. After more than a year of inactivity, HCMR was potentially faced with major changes and decisions. It was not until we recruited new staff and executive board members that I became fully confident that, once again, we were going to have an issue that would help students learn and appreciate advanced mathematics like the founder and former Editors had envisioned it would.

Since September when we started accepting applications for the different position until a few weeks ago when we were making final edits on the articles, members of the HCMR executive board and the staff have worked tirelessly to make this issue one of the strongest ones we have had. To this end, I would like to thank the Articles Editor, **Francois Greer**, the Features Editor, **Katherine Banks**, the Problems Editor, **Lucia Mocz** and their **staff** for all the hard work. I would also like to thank **Eric Larson** for taking up several roles and working on each flawlessly.

The executive board is thankful for **Professor Peter Kronheimer**, the **Harvard University Mathematics Department** and the staff at the **Student Organization Center at Hilles** for their continual advice and support. You continue to make the HCMR a success and we are glad to have your support year after year.

We also thank the HCMR **advisors** and **sponsors**, whose generous contributions have been a foundation for the journal's success.

I would also like to personally express my deepest gratitude to Editor Emeriti **Zach Abel**, **Scott D. Kominers** and **Ernest E. Fontes** for constantly being there to answer all the major and minor questions I had even after their graduation from Harvard College. Your invaluable expertise and guidance has inspired the entire staff.

Finally, the executive board would like to thank all our **writers** and **readers** in the United States and around the world. We are honored at all the feedback and contributions we have received so far. As always, please direct your comments, questions and submissions to hcmr@hcs.harvard.edu. As the board is in the process of discussing potential changes in the structure, web and print output, your feedback is more valuable than ever. For updates, check www.thehcmr.org/

Rediet Abebe
Editor-In-Chief, The HCMR

STUDENT ARTICLE

1 Artin's Conjecture

Christopher Policastro[†]
Massachusetts Institute of Technology '11
Cambridge, MA 02138
cpoli@mit.edu

Abstract

We survey Artin L-functions by providing the necessary background to describe Artin's conjecture. Having detailed the basic properties of Artin L-series, we show that they extend meromorphically to the plane, and discuss recent research on this continuation. We assume the reader has some knowledge of group representations, algebraic number theory, and complex analysis.

1.1 Introduction

Certain arithmetic properties of a number field are contained within holomorphic functions called L-series. These series are defined in some right half-plane and can be meromorphically continued to $\mathbb{C}$. The resulting L-functions share several analytic properties that can be used to characterize them.

The L-series for a number field take two rather different forms that can be reconciled in some cases. The abelian type were introduced by Erich Hecke, who constructed them from characters of ray class groups in an attempt to generalize the Dirichlet L-series

$$L(s,\chi) = \sum_{n \geq 1} \frac{\chi(n)}{n^s} \tag{1.1}$$

where $\chi : (\mathbb{Z}/m\mathbb{Z})^* \to \mathbb{C}^*$ is a homomorphism for some m. The nonabelian type owe to Emil Artin, who defined them by representations of Galois groups following the work of Takagi. Some years after he described the basic properties of these L-series, Richard Brauer proved that they admit a meromorphic continuation to the plane.

Artin famously conjectured that with few exceptions this continuation was holomorphic. Though progress continues to be made on the conjecture, it remains one of the most important outstanding problems in mathematics. The goal of this paper is to attract readers to this very active branch of analytic number theory.

The rest of the paper is as follows. Having recalled several facts about representations of finite groups and Hecke L-functions, we define incomplete Artin L-series and prove their functorial properties. Having shown that these series meromorphically extend to $\mathbb{C}$, we verify Artin's conjecture for a class of extensions, and briefly discuss works of Langlands and Selberg as they apply to the conjecture.

Before continuing let us fix some notation. Throughout we have k denote a number field, that is, a finite extension of $\mathbb{Q}$. Let $\mathcal{O}_k$ be the ring of integers, and for an ideal $\mathfrak{a}$ in $\mathcal{O}_k$, let $\mathfrak{N}(\mathfrak{a}) = \mathrm{Card}(\mathcal{O}_k/\mathfrak{a})$. For K/k a finite extension, and $\mathfrak{P}|\mathfrak{p}$ primes, we let $f_{\mathfrak{p}}^{\mathfrak{P}}$ and $e_{\mathfrak{p}}^{\mathfrak{P}}$ denote the residual degree and ramification index. By a place ν of k, we mean an equivalence class of nontrivial valuations of k. These are of three sorts: finite places corresponding to primes of k, real places corresponding to the real embeddings of k, and complex places corresponding to the pairs of conjugate complex embeddings of k.

[†]Christopher Policastro is a senior studying mathematics at the Massachusetts Institute of Technology. His mathematical interests lie somewhere between geometry and algebra. At the moment, he is leaning towards representation theory.

1.2 Representations of Finite Groups

In this section, we review results from the representation theory of finite groups that will be used in section 3. The basic definitions and properties that we omit here can be found in [3, Ch. 1,2].

1.2.1 Definitions

We recall that a *representation* of a finite group G on a finite dimensional $\mathbb{C}$-vector space V is a homomorphism $\rho : G \to \mathrm{GL}(V)$ of G into the group of invertible linear operators on V. This is equivalent to saying that V is a finite $\mathbb{C}[G]$-module.

The complex valued function $\chi(s) = \mathrm{Tr}(\rho(s))$ $s \in G$ is called the *character* of the representation ρ. We know that a character completely determines a representation, in the sense that two finite $\mathbb{C}[G]$-modules are isomorphic iff they have the same character. The character of a 1-dimensional representation is called *abelian*.

By Maschke's theorem, $\mathbb{C}[G]$ is semisimple. Therefore each finite $\mathbb{C}[G]$-module decomposes into a direct sum of *irreducible* representations. Characters of irreducible representations are themselves called irreducible; we denote the set of irreducible characters by $\widehat{G}$.

The group of *virtual characters* $K(G)$ is the free $\mathbb{Z}$-module spanned by the irreducible characters of G. Since a product of characters is itself a character, we see that $K(G)$ is also a ring.

A *class function* on G is a complex valued function that is constant on conjugacy classes. The space of class functions $F_{\mathbb{C}}(G)$ is a Hilbert space with respect to the Hermitian form

$$\langle \chi, \psi \rangle := \frac{1}{\mathrm{Card}(G)} \sum_{s \in G} \chi(s) \cdot \overline{\psi(s)}. \tag{1.2}$$

for $\chi, \psi \in F_{\mathbb{C}}(G)$. The elements of $\widehat{G}$ form an orthonormal basis of $F_{\mathbb{C}}(G)$.

It will be important for us to have a way of relating a representation (W, ρ) of a subgroup $H \subset G$ to a representation of G. We call the process of lifting ρ to G *induction*. It can be defined in two ways:

$$\mathrm{Ind}_H^G W := \mathbb{C}[G] \otimes_{\mathbb{C}[H]} W \quad \text{or} \quad \mathrm{Ind}_H^G W := \{ f : G \to W \mid f(\tau s) = \rho_\tau(f(s)),\ \forall \tau \in H \} \ .$$

Note that the action of $\mathbb{C}[G]$ on the second is given by $g(f) = f(\bullet g)$ for $g \in \mathbb{C}[G]$. The definitions are easily seen to be equal, and each will have its uses for us.

Given a representation (V, ρ) of G and a subgroup $H \subset G$, we obtain a representation $(\mathrm{Res}_H V, \rho|_H)$ by restricting ρ to H. Let χ be a character of G, and ψ a character of H. We recall that with respect to the formula in (2), Ind and Res act like adjoints, that is,

$$\left\langle \mathrm{Ind}_H^G \psi, \chi \right\rangle_G = \langle \psi, \mathrm{Res}_H \chi \rangle_H \ .$$

This equation is called Frobenius reciprocity.

1.2.2 Mackey's Theorem

Let G be a finite group with $H, K \subset G$ subgroups. Let $N \lhd K$, and (W, ρ) be a representation of H. For $V =: \mathrm{Ind}_H^G W$, we want to determine the restriction $\mathrm{Res}_K V$ of V to K, and in particular determine the subspace $(\mathrm{Res}_K V)^N$ of N-invariants. Note that $(\mathrm{Res}_K V)^N$ is a representation of K since N is normal in K.

Choose a set $K \setminus G/H$ of representatives for the (H, K) double cosets of G; this means G is the disjoint union of KsH for $s \in K \setminus G/H$. Let $H_s = sHs^{-1} \cap K$. Note that H_s is a subgroup of K, and the rule $\rho^s(x) := \rho(x)$ for $x \in H_s$ gives a homomorphism $\rho^s : H_s \to \mathrm{GL}(sW)$, where sW is taken in V. Inducing ρ^s to K for each s, we have the following result.

Proposition 1. $(\mathrm{Res}_K \mathrm{Ind}_H^G W)^N \cong \bigoplus_{s \in K \setminus G/H} (\mathrm{Ind}_{H_s}^K sW)^N$.

The proposition can be obtained as an immediate generalization of [6, Sec. 7.3].

1.2.3 Brauer's Theorem

As we will see in later sections, an L-series can be associated to abelian or nonabelian characters of certain groups. To make the connection between abelian and nonabelian L-series, we will need a way of relating arbitrary characters to abelian characters. A precise statement of this relation is given by Brauer's theorem.

Theorem 2. *Let G be a finite group. Each character of G is a $\mathbb{Z}$-linear combination of characters induced from abelian characters of subgroups.*

A version of the proof can be found in [6, Ch. 10].

1.3 Hecke L-functions

In this section, we review facts about a type of Hecke L-series. These functions have nice properties that apply to Artin L-series in certain cases. In particular, they extend with few exceptions to entire functions. This is exactly the kind of result we are after!

1.3.1 Definitions

Define a *modulus* $\mathfrak{m}$ as the formal product of a nonzero ideal $\mathfrak{m}_f$ of $\mathcal{O}_k$, and a set $\mathfrak{m}_\infty$ of real places of k. For $a, b \in \mathcal{O}_k$ such that (a),(b) are relatively prime to $\mathfrak{m}_f$, we take $a \equiv b (\mathrm{mod}^\times\ \mathfrak{m})$ to mean that $a/b \equiv 1 (\mathrm{mod}\ \mathfrak{m}_f)$ and $\sigma_\nu(a/b) > 0$ for each embedding σ_ν with $\nu \in \mathfrak{m}_\infty$. We can partially order moduli by the rule $\mathfrak{m} \leq \mathfrak{n}$ if $\mathfrak{n}_f \subset \mathfrak{m}_f$ and $\mathfrak{m}_\infty \subset \mathfrak{n}_\infty$.

Let $I_\mathfrak{m}^k$ be the group of fractional ideals of k relatively prime to $\mathfrak{m}_f$. We call the subgroup $P_\mathfrak{m}^k$ of principal ideals (a) such that $a \equiv 1 (\mathrm{mod}^\times\ \mathfrak{m})$ the *ray class* of $\mathfrak{m}$, and $Cl_\mathfrak{m}^k := I_\mathfrak{m}^k / P_\mathfrak{m}^k$ the *ray class group* of $\mathfrak{m}$. Ray class groups are known to be finite.

Example 3. If $k = \mathbb{Q}$ and $\mathfrak{m} = m \cdot \infty$ for $m > 0$, and we identify elements of $I_\mathfrak{m}^k$ with their positive generators, then we have a surjective map $I_\mathfrak{m}^k \to (\mathbb{Z}/m\mathbb{Z})^*$ with kernel given by those ideals (a) such that $a \equiv 1 (\mathrm{mod}\, m)$. This implies that the ray class group is $(\mathbb{Z}/m\mathbb{Z})^*$.

We call the quotient $Cl_\mathfrak{m}^k / H$, for some subgroup H, a *class group* of $\mathfrak{m}$. A Galois extension K/k is called a *class field* of $Cl_\mathfrak{m}^k / H$ if every prime which ramifies in K divides $\mathfrak{m}_f$, and the primes which split completely[1] are given by $P_\mathfrak{m}^k \cdot H$. A basic result tells us that for every class group of k, we can find a corresponding class field unique up to equivalence.

1.3.2 Hecke L-series

A *Dirichlet character* modulo $\mathfrak{m}$ is an abelian character χ of the ray class group $Cl_\mathfrak{m}^k$, which we extend to I^k by the rule $\chi(\mathfrak{a}) = 0$ for $\mathfrak{a}$ and $\mathfrak{m}$ not relatively prime. For $\mathfrak{m} \leq \mathfrak{n}$, the identity homomorphism on ideals gives a surjective homomorphism $Cl_\mathfrak{n}^k \to Cl_\mathfrak{m}^k$. This lets us think of characters modulo $\mathfrak{m}$ as characters modulo $\mathfrak{n}$. Moreover, for a Dirichlet character modulo $\mathfrak{m}$, we can find a smallest modulus $\mathfrak{m}_c$ dividing $\mathfrak{m}$ such that χ factors through $Cl_{\mathfrak{m}_c}^k$. We say that χ is *primitive* modulo $\mathfrak{m}_c$.

The Hecke L-series corresponding to a Dirichlet character modulo $\mathfrak{m}$ is defined as the sum

$$L(s, \chi) = \sum_\mathfrak{a} \frac{\chi(\mathfrak{a})}{\mathfrak{N}(\mathfrak{a})^s}$$

for $\mathrm{Re}(s) > 1$, where $\mathfrak{a}$ varies over integral ideals of k.

Proposition 4. *$L(s, \chi)$ is holomorphic in the domain $\mathrm{Re}(s) \geq 1 + \delta$ for all $\delta > 0$, and has the product decomposition $L(s, \chi) = \prod_\mathfrak{p} (1 - \chi(\mathfrak{p})\mathfrak{N}(\mathfrak{p})^{-s})^{-1}$ where $\mathfrak{p}$ ranges over all primes.*

Proof. Let

$$E(s) = \prod_\mathfrak{p} \frac{1}{1 - \chi(\mathfrak{p})\mathfrak{N}(\mathfrak{p})^{-s}}$$

[1] That is, the ramification index and residue class field degree are both 1.

for $\mathrm{Re}(s) \geq 1+\delta$ and $\mathfrak{p}$ such that $\chi(\mathfrak{p}) \neq 0$. Formally taking the logarithm gives

$$\log E(s) = \sum_{\mathfrak{p}} \sum_{n \geq 1} \frac{\chi(\mathfrak{p})^n}{n \mathfrak{N}(\mathfrak{p})^{ns}} .$$

Since χ is abelian, and $Cl_{\mathfrak{m}}^k$ is finite, we have $|\chi(\mathfrak{p})| = 1$. As $|\mathfrak{N}(\mathfrak{p})^s| = \mathfrak{N}(\mathfrak{p})^{\mathrm{Re}(s)} \geq p^{f_p^{\mathfrak{p}}(1+\delta)} \geq p^{1+\delta}$, and at most $[k:\mathbb{Q}]$ primes lie above p, we have that

$$\sum_{\mathfrak{p}} \sum_{n \geq 1} \frac{|\chi(\mathfrak{p})|^n}{n \mathfrak{N}(\mathfrak{p})^{n \,\mathrm{Re}(s)}} \leq \sum_{p} \sum_{n \geq 1} \frac{[k:\mathbb{Q}]}{n p^{n(1+\delta)}} = [k:\mathbb{Q}] \log \zeta(1+\delta) . \tag{1.3}$$

This bound does not depend on s. So we see that $\log E(s)$ converges absolutely and uniformly for $\mathrm{Re}(s) \geq 1+\delta$. Therefore $E(s)$ is holomorphic in this half plane, and we are left showing that $E(s) = L(s,\chi)$.

Expand the factors in $E(s)$ corresponding to the finitely many primes $\mathfrak{p}_1 \ldots \mathfrak{p}_r$ such that $\mathfrak{N}(\mathfrak{p}_i) \leq N$, and multiply them. This yields

$$\prod_{i=1}^{r} \frac{1}{1-\chi(\mathfrak{p}_i)\mathfrak{N}(\mathfrak{p}_i)^{-s}} = \prod_{i=1}^{r} \left(1 + \frac{\chi(\mathfrak{p}_i)}{\mathfrak{N}(\mathfrak{p}_i)^s} + \ldots\right) = \sum_{\mathfrak{N}(\mathfrak{a}) \leq N} \frac{\chi(\mathfrak{a})}{\mathfrak{N}(\mathfrak{a})^s} + \sum_{\mathfrak{N}(\mathfrak{a}) > N}{}' \frac{\chi(\mathfrak{a})}{\mathfrak{N}(\mathfrak{a})^s} \tag{1.4}$$

where the prime indicates that the second sum in the last equality is only over integral ideals $\mathfrak{a}$ that are divisible only by the primes $\mathfrak{p}_1 \ldots \mathfrak{p}_r$. By (4) we have

$$\left| \prod_{i=1}^{r} \frac{1}{1-\chi(\mathfrak{p}_i)\mathfrak{N}(\mathfrak{p}_i)^{-s}} - L(s,\chi) \right| \leq \sum_{\mathfrak{N}(\mathfrak{a}) > N} \frac{1}{\mathfrak{N}(\mathfrak{a})^{1+\delta}},$$

so we need to show that this last term tends to zero as $N \to \infty$ to complete the proof. Now

$$\sum_{\mathfrak{N}(\mathfrak{a}) \leq N} \frac{1}{\mathfrak{N}(\mathfrak{a})^{1+\delta}} \leq \sum_{\mathfrak{a}}{}' \frac{1}{\mathfrak{N}(\mathfrak{a})^{1+\delta}} = \prod_{i=1}^{r} \frac{1}{1-\mathfrak{N}(\mathfrak{p}_i)^{-(1+\delta)}},$$

where the second sum is over integral ideals only divisible by $\mathfrak{p}_1 \ldots \mathfrak{p}_r$. So using the bound in (3) for the case of χ trivial and $s = 1+\delta$, we see that $\sum_{\mathfrak{N}(\mathfrak{a}) \leq N} \mathfrak{N}(\mathfrak{a})^{-1-\delta}$ is monotone increasing and bounded above by $\zeta(1+\delta)^{[k:\mathbb{Q}]}$ as $N \to \infty$. Hence the tail $\sum_{\mathfrak{N}(\mathfrak{a}) > N} \mathfrak{N}(\mathfrak{a})^{-1-\delta}$ converges to zero as $N \to \infty$. □

From Example 3, we see that Hecke L-series do indeed generalize Dirichlet L-series. Though it requires some work, we can extend Hecke L-series to the plane in much the same way we extend Dirichlet L-series. We obtain the following result, due to Hecke.

Proposition 5. *$L(s,\chi)$ can be holomorphically continued to $\mathbb{C}$ for χ nontrivial. For χ trivial, it extends to a meromorphic function with poles at $s = 0, 1$.*

For a proof of this fact see [5, VII.8].

1.4 Artin L-functions

We are at last ready to describe Artin L-series. Since (1) is the Riemann zeta function for $m = 1$, the series $L(s, \chi_{\mathrm{triv}})$ for $\mathfrak{m} = \mathcal{O}_k$ has historically been denoted as $\zeta_k(s)$. When Artin introduced a L-series in [1] attached to nonabelian characters, he hoped to verify a conjecture of Dedekind concerning the poles of $\zeta_K(s)/\zeta_k(s)$ for K/k a Galois extension.

Assuming that K/k is a class field, Weber had shown that $\zeta_K(s)/\zeta_k(s)$ could be decomposed into a product of L-series for nontrivial characters of the corresponding class group. By Hecke's

result, these L-series could be holomorphically extended to the plane. Now Artin was aware of Takagi's work on class field theory, and realized, in particular, that every abelian extension of k is a class field. So knowing that $\zeta_K(s)/\zeta_k(s)$ is entire for abelian extensions, he wanted to show that the same was true of nonabelian extensions.

This prompted his research on L-functions, and led to further advances in class field theory.

1.4.1 Definition

Let K/k be a Galois extension, and $\rho : \mathrm{Gal}(K/k) \to \mathrm{GL}(V)$ a representation. For a prime $\mathfrak{p}$ in $\mathcal{O}_k$, choose $\mathfrak{P}|\mathfrak{p}$. Let $D_\mathfrak{P}$ and $I_\mathfrak{P}$ be the decomposition and inertia groups of $\mathfrak{P}$ over $\mathfrak{p}$, and choose an element $\mathrm{Fr}_\mathfrak{P} \in D_\mathfrak{P}$ that reduces to the Frobenius automorphism under the map $D_\mathfrak{P} \twoheadrightarrow \mathrm{Gal}((\mathcal{O}_K/\mathfrak{P})/(\mathcal{O}_k/\mathfrak{p}))$. By abuse of notation, we will call $\mathrm{Fr}_\mathfrak{P}$ a Frobenius element of $\mathfrak{P}$.

For $\mathrm{Re}(s) > 1$, we define the Artin L-series as

$$L(s, \rho, K/k) = \prod_\mathfrak{p} L_\mathfrak{p}(s, \rho, K/k) = \prod_\mathfrak{p} \frac{1}{\det\left(1 - \rho(\mathrm{Fr}_\mathfrak{P})\mathfrak{N}(\mathfrak{p})^{-s}\right)|_{V^{I_\mathfrak{P}}}} .$$

The notation means that we consider $\mathrm{Fr}_\mathfrak{P}$ as an element of the decomposition group $D_\mathfrak{P}$, and take its image under the representation $(\mathrm{Res}_{D_\mathfrak{P}} V)^{I_\mathfrak{P}}$. Note that restricting to the space of $I_\mathfrak{P}$ invariants yields a well-defined representation since $I_\mathfrak{P}$ is normal in $D_\mathfrak{P}$. We call $L_\mathfrak{p}(s, \rho, K/k)$ a *local factor*. Our first task is to show that this definition even makes sense.

Proposition 6. *The local factors in the L-series are well-defined, and do not depend on the isomorphism class of the representation of* $\mathrm{Gal}(K/k)$.

Proof. Let us show that $L_\mathfrak{p}(s, \rho, K/k)$ does not depend on our choice of Frobenius element.

For a given $\mathfrak{P}$, each Frobenius element is of the form $\mathrm{Fr}_\mathfrak{P}\,\tau$ for $\tau \in I_\mathfrak{P}$. As the action of $D_\mathfrak{P}$ is on $I_\mathfrak{P}$-invariants, we see that $\mathrm{Fr}_\mathfrak{P}\,\tau$ and $\mathrm{Fr}_\mathfrak{P}$ determine the same map. Suppose that instead of $\mathfrak{P}$, we picked $\mathfrak{Q}|\mathfrak{p}$. Choose $g \in \mathrm{Gal}(K/k)$ such that $\mathfrak{Q} = g(\mathfrak{P})$. We recall that $g\,\mathrm{Fr}_\mathfrak{P}\,g^{-1}$ is a Frobenius element of $\mathfrak{Q}$, and $I_\mathfrak{Q} = gI_\mathfrak{P}g^{-1}$. So $V^{I_\mathfrak{Q}} = gV^{I_\mathfrak{P}}$, and we would like to prove that

$$\det\left(1 - \rho(\mathrm{Fr}_\mathfrak{P})\mathfrak{N}(\mathfrak{p})^{-s}\,|_{V^{I_\mathfrak{P}}}\right) = \det\left(1 - \rho(g\,\mathrm{Fr}_\mathfrak{P}\,g^{-1})\mathfrak{N}(\mathfrak{p})^{-s}\,|_{gV^{I_\mathfrak{P}}}\right) .$$

This equality, however, is immediate, since the determinant does not change under conjugation.

Finally, we should check that the L-function does not depend on the isomorphism class of the representation. But this is clear from the definition. □

For ease of notation, we will start suppressing ρ in the expressions for local factors.

Proposition 7. *The Artin L-series is holomorphic in the domain* $\mathrm{Re}(s) \geq 1 + \delta$ *for all* $\delta > 0$.

Proof. Consider $L_\mathfrak{p}(s, \rho, K/k)$ for the N-dimensional representation (V, ρ). We can diagonalize the matrix corresponding to $\mathrm{Fr}_\mathfrak{P}$ to obtain

$$L_\mathfrak{p}(s, \rho, K/k) = \frac{1}{\det\left(1 - \mathrm{Fr}_\mathfrak{P}\,\mathfrak{N}(\mathfrak{p})^{-s}\right)|_{V^{I_\mathfrak{P}}}} = \prod_{i=1}^{d}(1 - \varepsilon_i\mathfrak{N}(\mathfrak{p})^{-s})^{-1}$$

where d is the dimension of the representation $(\mathrm{Res}_{D_\mathfrak{P}} V)^{I_\mathfrak{P}}$, and ε_i are the eigenvalues. So formally taking the logarithm, we have

$$\log L(s, \chi, K/k) = \sum_\mathfrak{p} \sum_{\varepsilon_i} \sum_{n \geq 1} \frac{\varepsilon_i^n}{n\mathfrak{N}(\mathfrak{p})^{ns}} . \tag{1.5}$$

Note that $|\varepsilon_i| = 1$ since $\mathrm{Fr}_\mathfrak{P}$ has finite order. So

$$\sum_\mathfrak{p} \left(\sum_{\varepsilon_i} \sum_{n \geq 1} \frac{|\varepsilon_i|^n}{n\mathfrak{N}(\mathfrak{p})^{n\,\mathrm{Re}(s)}} \right) \leq \sum_\mathfrak{p} N \sum_{n \geq 1} \frac{1}{n\mathfrak{N}(\mathfrak{p})^{n\,\mathrm{Re}(s)}} .$$

Therefore the result follows by comparison with the logarithm of $\zeta_k(s)$, and the argument in Proposition 4. □

Considering that the L-series depends on the representation of $\mathrm{Gal}(K/k)$ only up to isomorphism, we will sometimes denote it as $L(s, \chi, K/k)$ for the corresponding character χ. In fact, we can define the Artin L-series strictly in terms of a character. From (5), we have

$$\log L(s, \chi, K/k) = \sum_{\mathfrak{p}}{}' \sum_{\varepsilon_i} \sum_{n \geq 1} \frac{\varepsilon_i^n}{n\mathfrak{N}(\mathfrak{p})^{ns}} = \sum_{\mathfrak{p}}{}' \sum_{n \geq 1} \frac{\chi(\mathrm{Fr}_{\mathfrak{P}}^n)}{n\mathfrak{N}(\mathfrak{p})^{ns}} .$$

(Here the prime indicates that the terms are slightly different for the ramified primes; we omit the details.) Hence we can take $L(s, \chi, K/k)$ to be

$$L(s, \chi, K/k) = \exp\left(\sum_{\mathfrak{p}}{}' \sum_{n \geq 1} \frac{\chi(\mathrm{Fr}_{\mathfrak{P}}^n)}{n\mathfrak{N}(\mathfrak{p})^{ns}} \right). \tag{1.6}$$

Using this definition, we can make sense of Artin L-series for virtual character; namely, for $-\chi \in K(\mathrm{Gal}(K/k))$ with χ a character of $\mathrm{Gal}(K/k)$, we have

$$L(s, -\chi, K/k) = L(s, \chi, K/k)^{-1} .$$

Example 8. Consider the extension $\mathbb{Q}(i)/\mathbb{Q}$. While this example will not show us the real strength of Artin L-series, it will let us see some of their basic properties at work. So $\mathrm{Gal}(\mathbb{Q}(i)/\mathbb{Q}) = \{1, \varepsilon\}$, and we recall that primes congruent to $1 \pmod 4$ split, and $3 \pmod 4$ stay inert, while 2 ramifies. Hence

$$\mathrm{Fr}_{\mathfrak{P}} = \begin{cases} 1 & \text{if } \mathfrak{P}|p \text{ with } p \equiv 1 \pmod 4 \\ \varepsilon & \text{if } \mathfrak{P}|p \text{ with } p \equiv 3 \pmod 4 \end{cases} \tag{1.7}$$

Let $\rho : \mathrm{Gal}(\mathbb{Q}(i)/\mathbb{Q}) \to \mathrm{GL}_2(\mathbb{C})$ be the map given by $\rho(\varepsilon) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Since inertia is trivial for unramified primes, we see that

$$L_p(s, \rho, \mathbb{Q}(i)/\mathbb{Q}) = \begin{cases} (1 - p^{-s})^{-2} & \text{if } p \equiv 1 \pmod 4 \\ (1 - p^{-2s})^{-1} & \text{if } p \equiv 3 \pmod 4 \end{cases} .$$

For $p = 2$, we have $(1 + i)|2$, and $I_{1+i} = \mathrm{Gal}(\mathbb{Q}(i)/\mathbb{Q})$. The subspace of ε-invariants is spanned by the vector $(1, 1)^t$. Since ρ reduces to the trivial representation on this subspace, we conclude that $L_2(s, \rho, \mathbb{Q}(i)/\mathbb{Q}) = (1 - 2^{-s})^{-1}$ whether we choose 1 or ε for the Frobenius element. Hence

$$L(s, \rho, \mathbb{Q}(i)/\mathbb{Q}) = \frac{1}{1 - 2^{-s}} \prod_{p \equiv 1 \pmod 4} \frac{1}{(1 - p^{-s})^2} \prod_{p \equiv 3 \pmod 4} \frac{1}{1 - p^{-2s}} .$$

Since $(\mathbb{C}^2, \rho)$ is not irreducible, we can decompose $\mathbb{C}^2$ as $U \oplus W = \mathbb{C}\binom{1}{1} \oplus \mathbb{C}\binom{1}{-1}$ corresponding to the eigenspaces of ε. Since V is a direct sum of U and W, we can express each local factor as $L_p(s, \rho_U, \mathbb{Q}(i)/\mathbb{Q}) L_p(s, \rho_W, \mathbb{Q}(i)/\mathbb{Q})$. As ρ_U is trivial, it follows that $L(s, \rho_U, \mathbb{Q}(i)/\mathbb{Q}) = \zeta_{\mathbb{Q}}(s)$. On W, $\rho_W(\varepsilon)$ is scaling by -1. For $p \neq 2$, inertia is trivial and so by (7) we have

$$\rho_W(\mathrm{Fr}_{\mathfrak{P}}) = \begin{cases} 1 & \text{if } \mathfrak{P}|p \text{ with } p \equiv 1 \pmod 4 \\ -1 & \text{if } \mathfrak{P}|p \text{ with } p \equiv -1 \pmod 4 \end{cases} .$$

For $p = 2$, $W^{I_{1+i}} = 0$ which by convention yields $L_2(s, \rho_W, \mathbb{Q}(i)/\mathbb{Q}) = 1$. Therefore

$$L(s, \rho_W, \mathbb{Q}(i)/\mathbb{Q}) = \prod_{p \equiv 1 \pmod 4} \frac{1}{1 - p^{-s}} \prod_{p \equiv 3 \pmod 4} \frac{1}{1 + p^{-s}} .$$

We notice that the right hand side is the Hecke L-series for modulus $4 \cdot \infty$ and the nontrivial irreducible character of $(\mathbb{Z}/4\mathbb{Z})^*$, which we denote as χ_4. So we conclude that

$$L(s, \rho, \mathbb{Q}(i)/\mathbb{Q}) = \zeta_{\mathbb{Q}}(s) L(s, \chi_4) \, . \tag{1.8}$$

From this example, we can already see some of the important facts about Artin L-series. In particular, we guess that the L-series corresponding to the trivial representation is merely the zeta function of k; namely $L(s, \chi_{\text{triv}}, K/k) = \zeta_k(s)$. We will have more to say about such basic properties in the following section.

1.4.2 Functorial Properties

In this section, we prove three results that will be crucial to our understanding of Artin L-series. In particular, we learn how L-series corresponding to different fields relate to one another.

Proposition 9 (Additivity). *If χ and χ' are virtual characters of* $\mathrm{Gal}(K/k)$ *then* $L(s, \chi+\chi', K/k) = L(s, \chi, K/k) L(s, \chi', K/k)$.

Proof. This follows immediately from (6). We note that rearrangement is allowed since the series converges absolutely, as we saw in Proposition 7. □

Proposition 10 (Towers). *Let $k \subset K \subset L$ be such that K/k is Galois, and let χ be a character of* $\mathrm{Gal}(K/k)$. *Extend χ to a character χ' of* $\mathrm{Gal}(L/k)$. *It follows that* $L(s, \chi', L/k) = L(s, \chi, K/k)$.

Proof. Let χ correspond to a representation (V, ρ). Let $\mathfrak{P}' | \mathfrak{P} | \mathfrak{p}$ be prime ideals in $\mathcal{O}_L$, $\mathcal{O}_K$, and $\mathcal{O}_k$. $\mathrm{Gal}(L/k)$ acts on V according to the projection $\mathrm{Gal}(L/k) \to \mathrm{Gal}(K/k)$. This map induces surjective homomorphisms $D_{\mathfrak{P}'} \to D_{\mathfrak{P}}$ and $I_{\mathfrak{P}'} \to I_{\mathfrak{P}}$. So we obtain a surjective homomorphism $D_{\mathfrak{P}'}/I_{\mathfrak{P}'} \to D_{\mathfrak{P}}/I_{\mathfrak{P}}$ that sends $\mathrm{Fr}_{\mathfrak{P}'}$ to $\mathrm{Fr}_{\mathfrak{P}}$. Therefore the action of $\mathrm{Fr}_{\mathfrak{P}'}$ on $V^{I_{\mathfrak{P}'}}$ is the same as the action of $\mathrm{Fr}_{\mathfrak{P}}$ on $V^{I_{\mathfrak{P}}}$. This implies that

$$\det\left(1 - \mathrm{Fr}_{\mathfrak{P}'} \, \mathfrak{N}(\mathfrak{p})^{-s}\right) |_{V^{I_{\mathfrak{P}'}}} = \det\left(1 - \mathrm{Fr}_{\mathfrak{P}} \, \mathfrak{N}(\mathfrak{p})^{-s}\right) |_{V^{I_{\mathfrak{P}}}}$$

which gives the result. □

Lemma 11. *Let G be a finite group with $H \subset G$ a subgroup. If $N \lhd G$, and (W, ρ) is a representation of H, then*

$$(\mathrm{Ind}_H^G W)^N \cong \mathrm{Ind}_{H/(H \cap N)}^{G/N} W^{H \cap N} \, .$$

Proof. We will use our function definition of induction to obtain a natural isomorphism. We note that a H-function $f : G \to W$ is N invariant iff $f(x\tau) = f(x)$ for all $\tau \in N$, namely, f is constant on right, and so also left, cosets of N. This holds iff f is a $H/(H \cap N)$-function on G/N. Such a function takes values in $W^{H \cap N}$ since $\tau f(x) = f(\tau x) = f(x)$ for $\tau \in H \cap N$. □

Proposition 12 (Induced Representations). *Let L/k be a Galois extension. For K an intermediate field, and χ a character of* $\mathrm{Gal}(L/K)$, *one has*

$$L(s, \chi, L/K) = L(s, \chi', L/k)$$

where $\chi' := \mathrm{Ind}_{\mathrm{Gal}(L/K)}^{\mathrm{Gal}(L/k)} \chi$.

Proof. Let $G = \mathrm{Gal}(L/k)$ and $H = \mathrm{Gal}(L/K)$. Suppose that χ corresponds to a representation (W, ρ) of H, and take V to be $\mathrm{Ind}_H^G W$. Consider a prime $\mathfrak{p}$ in $\mathcal{O}_k$. Let $\mathfrak{q}_1, \ldots, \mathfrak{q}_r$ be the primes in $\mathcal{O}_K$ lying above $\mathfrak{p}$, and for each $\mathfrak{q}_i$, choose a prime $\mathfrak{P}_i$ in $\mathcal{O}_L$ dividing it. We want to show that

$$L_{\mathfrak{p}}(s, \chi', L/k) = \prod_{i=1}^{r} L_{\mathfrak{q}_i}(s, \chi, L/K) \tag{1.9}$$

and will proceed by reducing to the case $r = 1$, and then the case that $\mathfrak{p}$ is unramified.

Denote the decomposition and inertia groups of $\mathfrak{P}_i$ over $\mathfrak{p}$ by D_i and I_i. We see that $D_i' := H \cap D_i$ and $I_i' := H \cap I_i$ are the decomposition and inertia groups of $\mathfrak{P}_i$ over $\mathfrak{q}_i$. As G acts transitively on the primes lying above $\mathfrak{p}$, we can choose elements $\tau_i \in G$ such that $\tau_i^{-1}\mathfrak{P}_1 = \mathfrak{P}_i$. Since H acts transitively on the primes dividing each $\mathfrak{q}_i$, we note that the set $\{\tau_i \mid 1 \leq i \leq r\}$ is a system of representatives for the double cosets $D_1 \setminus G/H$.

We recall that $D_i = \tau_i^{-1} D_1 \tau_i$, $I_i = \tau_i^{-1} I_1 \tau_i$, and $\tau_i^{-1} \operatorname{Fr}_{\mathfrak{P}_1} \tau_i = \operatorname{Fr}_{\mathfrak{P}_i}$. For f_i the residual degree of $\mathfrak{q}_i$ over $\mathfrak{p}$, we know that $\mathfrak{N}(\mathfrak{q}_i) = \mathfrak{N}(\mathfrak{p})^{f_i}$, and $\operatorname{Fr}_{\mathfrak{P}_i}^{f_i}$ is a Frobenius element of $\mathfrak{P}_i$ over $\mathfrak{q}_i$.

To reduce to the case of $r = 1$, we will use Mackey's theorem. We can take

$$L_{\mathfrak{p}}(s, \chi', L/k) = \frac{1}{\det\left(1 - \operatorname{Fr}_{\mathfrak{P}_1} \mathfrak{N}(\mathfrak{p})^{-s}\right) |_{V^{I_1}}} .$$

Since $\operatorname{Fr}_{\mathfrak{P}_1} \in D_1$, this can be rewritten as

$$L_{\mathfrak{p}}(s, \chi', L/k) = \left(\det\left(1 - \operatorname{Fr}_{\mathfrak{P}_1} \mathfrak{N}(\mathfrak{p})^{-s}\right) |_{(\operatorname{Res}_{D_1} \operatorname{Ind}_H^G W)^{I_1}}\right)^{-1} .$$

By Proposition 6, we know that local factors are equivalent under isomorphism. So by Proposition 1, we have

$$\begin{aligned} L_{\mathfrak{p}}(s, \chi', L/k) &= \left(\det(1 - \operatorname{Fr}_{\mathfrak{P}_1} \mathfrak{N}(\mathfrak{p})^{-s} \mid \bigoplus_{i=1}^{r} (\operatorname{Ind}_{D_1 \cap \tau_i H \tau_i^{-1}}^{D_1} \tau_i W)^{I_1})\right)^{-1} \\ &= \prod_{i=1}^{r} (\det(1 - \operatorname{Fr}_{\mathfrak{P}_1} \mathfrak{N}(\mathfrak{p})^{-s} \mid (\operatorname{Ind}_{D_1 \cap \tau_i H \tau_i^{-1}}^{D_1} \tau_i W)^{I_1}))^{-1} . \end{aligned}$$

We can conjugate each factor by τ_i^{-1} to obtain

$$\begin{aligned} L_{\mathfrak{p}}(s, \chi', L/k) &= \prod_{i=1}^{r} (\det(1 - \tau_i^{-1} \operatorname{Fr}_{\mathfrak{P}_1} \tau_i \mathfrak{N}(\mathfrak{p})^{-s} \mid (\operatorname{Ind}_{\tau_i^{-1} D_1 \tau_i \cap H}^{\tau_i^{-1} D_1 \tau_i} W)^{\tau_i^{-1} I_1 \tau_i}))^{-1} \\ &= \prod_{i=1}^{r} (\det(1 - \operatorname{Fr}_{\mathfrak{P}_i} \mathfrak{N}(\mathfrak{p})^{-s} \mid (\operatorname{Ind}_{D_i'}^{D_i} W)^{I_i}))^{-1}. \end{aligned}$$

Therefore (9) can be rewritten as

$$\begin{aligned} \prod_{i=1}^{r} (\det(1 - \operatorname{Fr}_{\mathfrak{P}_i} \mathfrak{N}(\mathfrak{p})^{-s} \mid (\operatorname{Ind}_{D_i'}^{D_i} W)^{I_i}))^{-1} &= \prod_{i=1}^{r} L_{\mathfrak{q}_i}(s, \chi, L/K) \\ &= \prod_{i=1}^{r} \left(\det(1 - \operatorname{Fr}_{\mathfrak{P}_i}^{f_i} \mathfrak{N}(\mathfrak{p})^{-f_i s}) |_{W^{I_i'}}\right)^{-1} . \end{aligned}$$

This equation will follow from equality of the i^{th} factors on either side. So without loss of generality, assume that $r = 1$. We are left showing that

$$(\det(1 - \operatorname{Fr}_{\mathfrak{P}_i} \mathfrak{N}(\mathfrak{p})^{-s} \mid (\operatorname{Ind}_{D_i'}^{D_i} W)^{I_i}))^{-1} = (\det(1 - \operatorname{Fr}_{\mathfrak{P}_i}^{f_i} \mathfrak{N}(\mathfrak{p})^{-f_i s} \mid W^{I_i'}))^{-1}$$

for all i. By Lemma 11, this can be rewritten as

$$(\det(1 - \operatorname{Fr}_{\mathfrak{P}_i} \mathfrak{N}(\mathfrak{p})^{-s} \mid (\operatorname{Ind}_{D_i'/I_i'}^{D_i/I_i} W^{I_i'}))^{-1} = (\det(1 - \operatorname{Fr}_{\mathfrak{P}_i}^{f_i} \mathfrak{N}(\mathfrak{p})^{-f_i s} \mid W^{I_i'}))^{-1} \quad (1.10)$$

If we can convince ourselves that equation (10) corresponds to the tower $L^{D_i} \subset L^{D_i' I_i} \subset L^{I_i}$, then we can further assume that $\mathfrak{p}$ is unramified.

In this case, we have $\mathrm{Gal}(L^{I_i}/L^{D_i}) \cong D_i/I_i$, and $\mathrm{Gal}(L^{I_i}/L^{D_i'I_i}) \cong D_i'I_i/I_i \cong D_i'/I_i'$. The latter isomorphisms let us make sense of $W^{I_i'}$ as a representation of $\mathrm{Gal}(L^{I_i}/L^{D_i'I_i})$. Let

$$\mathfrak{Q}_D := \mathcal{O}_{L^{D_i}} \cap \mathfrak{P}_i, \qquad \mathfrak{Q}_{D'I} := \mathcal{O}_{L^{D_i'I_i}} \cap \mathfrak{P}_i, \qquad \mathfrak{Q}_I := \mathcal{O}_{L^{I_i}} \cap \mathfrak{P}_i\,.$$

Since

$$f_{\mathfrak{Q}_D}^{\mathfrak{Q}_I} = [L^{I_i} : L^{D_i}] = \mathrm{Card}(D_i/I_i) = f_{\mathfrak{p}}^{\mathfrak{P}_i}, \quad f_{\mathfrak{Q}_{D'I}}^{\mathfrak{Q}_I} = [L^{I_i} : L^{D_i'I_i}] = \mathrm{Card}(D_i'/I_i') = f_{\mathfrak{q}_i}^{\mathfrak{P}_i}$$

we find that $f_{\mathfrak{Q}_D}^{\mathfrak{Q}_{D'I}} = f_i$. Note that $\mathfrak{N}(\mathfrak{p}) = \mathfrak{N}(\mathfrak{Q}_D)$, and that the Frobenius element of $\mathfrak{Q}_I$ over $\mathfrak{Q}_D$ is $\mathrm{Fr}_{\mathfrak{P}_i}$. We conclude that for the sake of proving (10), we can take $\mathfrak{p}$, $\mathfrak{q}_i$, and $\mathfrak{P}_i$ to be $\mathfrak{Q}_D$, $\mathfrak{Q}_{D'I}$, and $\mathfrak{Q}_I$.

Keeping the initial notation, we can assume that $\mathfrak{P}$ is the only prime in $\mathcal{O}_L$ dividing $\mathfrak{p}$, and that it is unramified. We have that $G = \mathrm{Gal}(L/k)$ is generated by $\mathrm{Fr}_{\mathfrak{P}}$, and $H = \mathrm{Gal}(L/K)$ is generated by $\mathrm{Fr}_{\mathfrak{P}}^f$ for $f := f_{\mathfrak{p}}^{\mathfrak{q}} = [G : H]$. Using our extension of scalars definition of induction, we have that

$$V = \bigoplus_{i=0}^{f-1} \mathrm{Fr}_{\mathfrak{P}}^i W\,.$$

Let A be the matrix of $\mathrm{Fr}_{\mathfrak{P}}^f$ with respect to the basis $w_1, \ldots, w_d$ of W. If I is the $d \times d$ unit matrix, then

$$\begin{pmatrix} 0 & \cdots & 0 & A \\ I & \cdots & 0 & 0 \\ \vdots & \cdots & \vdots & \vdots \\ 0 & \cdots & I & 0 \end{pmatrix}$$

is the matrix of $\mathrm{Fr}_{\mathfrak{P}}$ with respect to the basis $\left\{\mathrm{Fr}_{\mathfrak{P}}^i w_j\right\}$ of V. This yields

$$\det(1 - \mathrm{Fr}_{\mathfrak{P}}\, \mathfrak{N}(\mathfrak{p})^{-s} \mid V) = \det \begin{pmatrix} I & \cdots & 0 & -\mathfrak{N}(\mathfrak{p})^{-s} A \\ -\mathfrak{N}(\mathfrak{p})^{-s} I & \cdots & 0 & 0 \\ \vdots & \cdots & \vdots & \vdots \\ 0 & \cdots & -\mathfrak{N}(\mathfrak{p})^{-s} I & I \end{pmatrix}. \quad (1.11)$$

Now multiply the first row by $\mathfrak{N}(\mathfrak{p})^{-s}$ and add it to the second, and then multiply the second row by $\mathfrak{N}(\mathfrak{p})^{-s}$ and add it to the third, and so on. Continuing in this way, we see that the determinant of the matrix in (11) is $\det\left(1 - \mathrm{Fr}_{\mathfrak{P}}^f\, \mathfrak{N}(\mathfrak{p})^{-sf}\right)|_W$. □

Consider the character χ'_{triv} of $\mathrm{Gal}(K/k)$ induced from the trivial character χ_{triv} of the subgroup $\{1\}$. By Frobenius reciprocity we have

$$\left\langle \chi'_{\mathrm{triv}}, \psi \right\rangle_{\mathrm{Gal}(K/k)} = \left\langle \chi_{\mathrm{triv}}, \mathrm{Res}_{\{1\}}\, \psi \right\rangle_{\{1\}} = \psi(1)$$

for each $\psi \in \widehat{\mathrm{Gal}}(K/k)$. Hence $\chi'_{\mathrm{triv}} = \sum_{\psi \in \widehat{\mathrm{Gal}}(K/k)} \psi(1) \cdot \psi$. So by Propositions 9 and 12, we obtain the following result.

Corollary 13. $\zeta_K(s) = \zeta_k(s) \prod_{\psi \in \widehat{\mathrm{Gal}}(K/k) \setminus \{1\}} L(s, \psi, K/k)^{\psi(1)}$.

Given that each $L(s, \psi, K/k)$ could be extended to an entire function, we would have the sought after generalization of Weber's result on abelian extensions! This remains conjectural, however, as we will soon discuss.

1.4.3 Abelian Characters

Having described some of the basic properties of Artin L-series, we want to know their relation to Hecke L-series. In particular, we might wonder whether an Artin L-series attached to an abelian character gives us a Hecke L-series. This seemed plausible to Artin, who *needed to relate his* factorization of $\zeta_K(s)$ (cf. Corollary 13) to Weber's factorization of $\zeta_K(s)$ in the case that K/k is abelian.

Proposition 14 *(Artin Reciprocity).* [illegible] *determines a surjective morphism* [illegible] *Takagi's construction* $P_{\mathfrak{f}}^k$ *in its kernel.*

For a discussion of this "dual form" of Artin Reciprocity see [7]. Let χ be an abelian character of $\mathrm{Gal}(K/k)$. Composing with the Artin symbol, and noting that this map factors through $I_{\mathfrak{f}}^k/P_{\mathfrak{f}}^k$, we obtain a Dirichlet character modulo $\mathfrak{f}$ of the ray class group. This allows us to give a precise statement of the relation of abelian L-series to nonabelian L-series.

Proposition 15. *Let K/k be an abelian extension with conductor $\mathfrak{f}$. Let $\chi \neq \chi_{triv}$ be an abelian character of* $\mathrm{Gal}(K/k)$, *and χ' the corresponding character of $Cl_{\mathfrak{f}}^k$. For $S := \{\mathfrak{p}|\mathfrak{f} \mid \chi(I_{\mathfrak{P}}) = 1\}$, it follows that*

$$L(s,\chi,K/k) = L(s,\chi') \prod_{\mathfrak{p}\in S} \frac{1}{1-\chi(\mathrm{Fr}_{\mathfrak{P}})\mathfrak{N}(\mathfrak{p})^{-s}} .$$

Proof. Since $\mathfrak{f}$ is the conductor of K/k, we recall that $\mathfrak{p}|\mathfrak{f}$ iff $\mathfrak{p}$ is ramified. If $\mathfrak{p}|\mathfrak{f}$ and $\chi(I_{\mathfrak{P}}) \neq 1$, then $\mathbb{C}^{I_{\mathfrak{P}}} = \{0\}$. So the corresponding local factor is 1. If $\mathfrak{p}|\mathfrak{f}$ and $\chi(I_{\mathfrak{P}}) = 1$, then

$$L_{\mathfrak{p}}(s,\chi,K/k) = \frac{1}{1-\chi(\mathrm{Fr}_{\mathfrak{P}})\mathfrak{N}(\mathfrak{p})^{-s}} .$$

Hence

$$L(s,\chi,K/k) = \prod_{\mathfrak{p}\nmid\mathfrak{f}} \frac{1}{1-\chi(\mathrm{Fr}_{\mathfrak{P}})\mathfrak{N}(\mathfrak{p})^{-s}} \prod_{\mathfrak{p}\in S} \frac{1}{1-\chi(\mathrm{Fr}_{\mathfrak{P}})\mathfrak{N}(\mathfrak{p})^{-s}} .$$

We recall that

$$L(s,\chi') = \prod_{\mathfrak{p}\nmid\mathfrak{f}} \frac{1}{1-\chi'(\mathfrak{p})\mathfrak{N}(\mathfrak{p})^{-s}} .$$

By construction $\chi'(\mathfrak{p}) = \chi\left(\left(\frac{K/k}{\mathfrak{p}}\right)\right)$. Therefore $\chi'(\mathfrak{p}) = \chi(\mathrm{Fr}_{\mathfrak{P}})$. This gives the result. □

So if χ is injective, then $S = \emptyset$, and we conclude that $L(s, \chi, K/k) = L(s, \chi')$. In the case that $\chi = \chi_{\text{triv}}$, then χ' is the trivial character modulo $\mathfrak{f}$, and we see that

$$\zeta_K(s) = L(s, \chi') \prod_{\mathfrak{p}|\mathfrak{f}} \frac{1}{1 - \mathfrak{N}(\mathfrak{p})^{-s}} .$$

Example 16. Let us see how the previous results apply to Example 8. By Corollary 13, we can replace $L(s, \rho, \mathbb{Q}(i)/\mathbb{Q})$ by $\zeta_{\mathbb{Q}(i)}(s)$ in (8) to obtain

$$\zeta_{\mathbb{Q}(i)}(s) = \zeta_{\mathbb{Q}}(s) L(s, \chi_4) . \tag{1.12}$$

Since the discriminant of $\mathbb{Q}(i)/\mathbb{Q}$ is 4, we know that the finite part of the conductor is either 2 or 4. We know that a prime $p > 0$ splits completely iff $p \equiv 1 (\mathrm{mod}\, 4)$. So $\mathfrak{f} = 4 \cdot \infty$ with $Cl_{4\cdot\infty}^{\mathbb{Q}} \cong (\mathbb{Z}/4\mathbb{Z})^*$ by Example 3. Hence Proposition 15 explains why χ_4 appears in expression (12).

1.4.4 Meromorphic Continuation.

denominators, we conclude that [illegible] p

1.5 Artin's Conjecture

Recall that the starting point of Artin's investigation of nonabelian L-series was the question of whether $\zeta_K(s)/\zeta_k(s)$ was an entire function for K/k a nonabelian Galois extension. By Corollary 13 and Theorem 17, we see that $\zeta_K(s)/\zeta_k(s)$ extends to a meromorphic function on $\mathbb{C}$. The results discussed do not, however, let us conclude that this extended function lacks poles. The difficulty is that, in the notation of Theorem 17, Brauer's theorem does not tell us which n_i are positive, or which ψ_i are nontrivial. Artin believed though that his nonabelian L-series could be analytically continued, given that the character χ did not contain any copies of χ_{triv}; namely, he made the following conjecture.

Artin's Conjecture. *For K/k a Galois extension, and χ a nontrivial irreducible character, the Artin L-series $L(s, \chi, K/k)$ extends to an entire function.*

Though progress has been made on this conjecture, it remains an open question. Before discussing some present work on this problem, let us note a class of Galois groups for which it is true. Suppose that $\mathrm{Gal}(K/k)$ is a *monomial* group, that is, each irreducible character χ of $\mathrm{Gal}(K/k)$ is induced from an abelian character ψ' of some subgroup H. This would include, for instance, the case that $\mathrm{Gal}(K/k)$ is a p-group for some prime p.

1.4.3 Abelian Characters

Having described some of the basic properties of Artin L-series, we want to know their relation to Hecke L-series. In particular, we might wonder whether an Artin L-series attached to an abelian character gives us a Hecke L-series. This seemed plausible to Artin, who needed to relate his factorization of $\zeta_K(s)$ (cf. Corollary 13) to Weber's factorization of $\zeta_K(s)$ in the case that K/k is abelian.

In this case, Artin knew that K is a class field for some class group $Cl_{\mathfrak{m}}^k/H$. Moreover, he knew that there exists a surjective homomorphism $Cl_{\mathfrak{m}}^k \twoheadrightarrow \mathrm{Gal}(K/k)$. However, Takagi's construction of this map was not explicit. If Artin could show that it took a canonical form, then he could prove that the Hecke L-series discussed in section 2 were subsumed under his L-series.

More specifically, let K/k be an abelian extension of number fields. There exists a certain modulus $\mathfrak{f}$ called the *conductor* with the property that a prime $\mathfrak{p}$ is unramified iff $\mathfrak{p} \nmid \mathfrak{f}$. Since the extension is abelian, each unramified prime has a unique Frobenius element, which we denote as $(\frac{K/k}{\mathfrak{p}})$. We obtain a canonical homomorphism

$$\left(\frac{K/k}{\cdot}\right) : I_{\mathfrak{f}}^k \to \mathrm{Gal}(K/k)$$

by setting

$$\left(\frac{K/k}{\mathfrak{a}}\right) = \prod_{\mathfrak{p}} \left(\frac{K/k}{\mathfrak{p}}\right)^{n_{\mathfrak{p}}}$$

for $\mathfrak{a} = \prod_{\mathfrak{p}} \mathfrak{p}^{n_{\mathfrak{p}}}$. We call $\left(\frac{K/k}{\mathfrak{a}}\right)$ the *Artin symbol.* Artin's rephrasing of Takagi's work is given by the following celebrated result.

Proposition 14 (Artin Reciprocity). $\left(\frac{K/k}{\cdot}\right)$ *determines a surjective homomorphism that contains* $P_{\mathfrak{f}}^k$ *in its kernel.*

For a discussion of this "dual form" of Artin Reciprocity see [7]. Let χ be an abelian character of $\mathrm{Gal}(K/k)$. Composing with the Artin symbol, and noting that this map factors through $I_{\mathfrak{f}}^k/P_{\mathfrak{f}}^k$, we obtain a Dirichlet character modulo $\mathfrak{f}$ of the ray class group. This allows us to give a precise statement of the relation of abelian L-series to nonabelian L-series.

Proposition 15. *Let K/k be an abelian extension with conductor $\mathfrak{f}$. Let $\chi \neq \chi_{triv}$ be an abelian character of* $\mathrm{Gal}(K/k)$*, and χ' the corresponding character of $Cl_{\mathfrak{f}}^k$. For $S := \{\mathfrak{p} | \mathfrak{f} \mid \chi(I_{\mathfrak{P}}) = 1\}$, it follows that*

$$L(s, \chi, K/k) = L(s, \chi') \prod_{\mathfrak{p} \in S} \frac{1}{1 - \chi(\mathrm{Fr}_{\mathfrak{P}})\mathfrak{N}(\mathfrak{p})^{-s}} \cdot$$

Proof. Since $\mathfrak{f}$ is the conductor of K/k, we recall that $\mathfrak{p} | \mathfrak{f}$ iff $\mathfrak{p}$ is ramified. If $\mathfrak{p} | \mathfrak{f}$ and $\chi(I_{\mathfrak{P}}) \neq 1$, then $\mathbb{C}^{I_{\mathfrak{P}}} = \{0\}$. So the corresponding local factor is 1. If $\mathfrak{p} | \mathfrak{f}$ and $\chi(I_{\mathfrak{P}}) = 1$, then

$$L_{\mathfrak{p}}(s, \chi, K/k) = \frac{1}{1 - \chi(\mathrm{Fr}_{\mathfrak{P}})\mathfrak{N}(\mathfrak{p})^{-s}} \cdot$$

Hence

$$L(s, \chi, K/k) = \prod_{\mathfrak{p} \nmid \mathfrak{f}} \frac{1}{1 - \chi(\mathrm{Fr}_{\mathfrak{P}})\mathfrak{N}(\mathfrak{p})^{-s}} \prod_{\mathfrak{p} \in S} \frac{1}{1 - \chi(\mathrm{Fr}_{\mathfrak{P}})\mathfrak{N}(\mathfrak{p})^{-s}} \cdot$$

We recall that

$$L(s, \chi') = \prod_{\mathfrak{p} \nmid \mathfrak{f}} \frac{1}{1 - \chi'(\mathfrak{p})\mathfrak{N}(\mathfrak{p})^{-s}} \cdot$$

By construction $\chi'(\mathfrak{p}) = \chi\left(\left(\frac{K/k}{\mathfrak{p}}\right)\right)$. Therefore $\chi'(\mathfrak{p}) = \chi(\mathrm{Fr}_{\mathfrak{P}})$. This gives the result. □

So if χ is injective, then $S = \emptyset$, and we conclude that $L(s, \chi, K/k) = L(s, \chi')$. In the case that $\chi = \chi_{\text{triv}}$, then χ' is the trivial character modulo $\mathfrak{f}$, and we see that

$$\zeta_K(s) = L(s, \chi') \prod_{\mathfrak{p}|\mathfrak{f}} \frac{1}{1 - \mathfrak{N}(\mathfrak{p})^{-s}} .$$

Example 16. Let us see how the previous results apply to Example 8. By Corollary 13, we can replace $L(s, \rho, \mathbb{Q}(i)/\mathbb{Q})$ by $\zeta_{\mathbb{Q}(i)}(s)$ in (8) to obtain

$$\zeta_{\mathbb{Q}(i)}(s) = \zeta_{\mathbb{Q}}(s) L(s, \chi_4) . \tag{1.12}$$

Since the discriminant of $\mathbb{Q}(i)/\mathbb{Q}$ is 4, we know that the finite part of the conductor is either 2 or 4. We know that a prime $p > 0$ splits completely iff $p \equiv 1 (\mathrm{mod}\, 4)$. So $\mathfrak{f} = 4 \cdot \infty$ with $Cl^{\mathbb{Q}}_{4,\infty} \cong (\mathbb{Z}/4\mathbb{Z})^*$ by Example 3. Hence Proposition 15 explains why χ_4 appears in expression (12).

1.4.4 Meromorphic Continuation.

Having connected Artin L-series to Hecke L-series, we can use Proposition 5 to meromorphically extend Artin L-series to the plane.

Theorem 17. *Let L/k be a Galois extension, and χ a character of* $\mathrm{Gal}(L/k)$. *The Artin L-series $L(s, \chi, L/k)$ admits a meromorphic continuation to* $\mathbb{C}$.

Proof. By Theorem 2, we can express χ as $\chi = \sum_{i=1}^{m} n_i \cdot \mathrm{Ind}_{H_i}^{G} \psi_i$ for $n_i \in \mathbb{Z}$, where $H_i \subset G$ are subgroups, and ψ_i is abelian. By Proposition 9 we obtain

$$L(s, \chi, L/k) = \prod_{i=1}^{m} L(s, \mathrm{Ind}_{H_i}^{G} \psi_i, L/k)^{n_i} .$$

By Proposition 12, we have $L(s, \mathrm{Ind}_{H_i}^{G} \psi_i, L/k) = L(s, \psi_i, L/K_i)$ where K_i is an intermediate field such that $\mathrm{Gal}(L/K_i) \cong H_i$. By Proposition 15 and Proposition 5, we know that $L(s, \psi_i, L/K_i)$ can be meromorphically continued to $\mathbb{C}$. This gives the result. □

Phew! Artin was unable to prove Theorem 17, though he made progress on it by showing a weaker version of Brauer's theorem. He proved that a character of a finite group can be expressed as a $\mathbb{Q}$-linear combination of characters induced from abelian characters of subgroups. So clearing denominators, he concluded that a sufficiently large power of the L-series admitted a continuation.

1.5 Artin's Conjecture

Recall that the starting point of Artin's investigation of nonabelian L-series was the question of whether $\zeta_K(s)/\zeta_k(s)$ was an entire function for K/k a nonabelian Galois extension. By Corollary 13 and Theorem 17, we see that $\zeta_K(s)/\zeta_k(s)$ extends to a meromorphic function on $\mathbb{C}$. The results discussed do not, however, let us conclude that this extended function lacks poles. The difficulty is that, in the notation of Theorem 17, Brauer's theorem does not tell us which n_i are positive, or which ψ_i are nontrivial. Artin believed though that his nonabelian L-series could be analytically continued, given that the character χ did not contain any copies of χ_{triv}; namely, he made the following conjecture.

Artin's Conjecture. *For K/k a Galois extension, and χ a nontrivial irreducible character, the Artin L-series $L(s, \chi, K/k)$ extends to an entire function.*

Though progress has been made on this conjecture, it remains an open question. Before discussing some present work on this problem, let us note a class of Galois groups for which it is true. Suppose that $\mathrm{Gal}(K/k)$ is a *monomial* group, that is, each irreducible character χ of $\mathrm{Gal}(K/k)$ is induced from an abelian character ψ' of some subgroup H. This would include, for instance, the case that $\mathrm{Gal}(K/k)$ is a p-group for some prime p.

From Frobenius reciprocity, we see that if χ is nontrivial, then ψ' is nontrivial. By Proposition 12, we know that that $L(s, \chi, K/k) = L(s, \psi', K/K^H)$. If $N \subset H$ is the kernel of ψ', then by Proposition 10, we have that $L(s, \psi', K/K^H) = L(s, \psi, K^N/K^H)$ where $\psi : H/N \hookrightarrow \mathbb{C}^*$. So by the remark after Proposition 15, and Proposition 5, we conclude that $L(s, \psi, K^N/K^H)$ analytically extends to $\mathbb{C}$. Therefore Artin's conjecture holds for monomial Galois groups.

What stinks is that not every group is monomial.

1.5.1 Langlands program

While Artin was able to prove many important analytic properties of his nonabelian L-series, and connect them to abelian L-series using his reciprocity theorem, he was unable to give the appropriate n-dimensional analogues of Dirichlet characters and L-functions. Although at the time, Hecke was researching such functions in the case of $n = 2$, it remained for Robert Langlands many years later to see a connection, and provide some precise statements. His vast set of conjectures offer not only a solution to Artin's conjecture, but a synthesis of many of the classical ideas in number theory.

We will very roughly describe Langlands' insight. For a more detailed survey the reader should see [2, Sec. III,IV] and [4, Sec. 7]. Given a place ν of k, let k_ν denote the corresponding completion with valuation ring $\mathcal{O}_\nu$. Let $G_n(\mathbf{A})$ be the subgroup of $\prod_\nu \mathrm{GL}_n(k_\nu)$ formed by tuples (g_ν) such that $g_\nu \in \mathrm{GL}_n(\mathcal{O}_\nu)$ for all but finitely many ν. Suppose π is an irreducible unitary representation of $G_n(\mathbf{A})$ in some Hilbert space H_π. By defining local factors for almost all primes, Langlands described an L-series $L(s, \pi)$ given by their product. Jacquet and Langlands then proved that for arbitrary π, local factors could be added so that the product could be taken over all primes. Moreover, if π takes a certain form, they were able to show that $L(s, \pi)$ extends to an entire function, unless $n = 1$ and π is trivial. Motivated by a search for a sort of converse to this result, Langlands made the following conjecture.

Langlands' Reciprocity Conjecture. *Let K/k be a Galois extension, and $\sigma : \mathrm{Gal}(K/k) \to \mathrm{GL}_n(\mathbb{C})$ an irreducible n-dimensional representation. There exists a representation π_σ of $G_n(\mathbf{A})$ such that $L(s, \pi_\sigma) = L(s, \sigma, K/k)$.*

When $n = 1$ and K/k is abelian, this conjecture reduces to Artin's reciprocity theorem. For arbitrary n, the truth of this result would imply Artin's conjecture.

As we will see in the next section, there appear to be other ways of arriving at Artin's conjecture. But given the scope of the Langlands program, this route would be a momentous achievement for number theory.

1.5.2 Selberg conjectures

In the course of this paper, we have seen several different constructions of L-series. The similarities of these constructions lead us to wonder whether we could study a broadly defined family of L-functions, rather than a single construction. These sorts of concerns led Alte Selberg to axiomatically define a class $\mathcal{S}$ of L-functions. Elements $F(s) \in S$ are complex valued functions of a complex variable that satisfy several of the properties we have discussed; they should be representable as a series for $\mathrm{Re}(s) > 1$, but also have an expression as a product, and they should meromorphically extend to the plane.

Since $\mathcal{S}$ is multiplicatively closed, Selberg wanted some notion of factorization and irreducibility. He called a function $F \in \mathcal{S}$ primitive if the equation $F = F_1 F_2$ for $F_1, F_2 \in \mathcal{S}$ implies either $F = F_1$ of $F = F_2$. It can be shown that each element of $\mathcal{S}$ factors into a product of primitive functions. The uniqueness of this factorization would be among several consequences of two conjectures made by Selberg. For statements, we refer the reader to [4, Sec. 1].

Using the Chebotarev density theorem, and the basic properties of Artin L-series developed here, it is easily shown that unique factorization implies Artin's conjecture. For a proof of this fact, and a more detailed discussion of the relation of Selberg's conjectures to Artin's and Langlands' conjectures see [4].

Acknowledgements

The author would like to thank Keith Conrad for suggesting several revisions to this paper, and also Carl Erickson and Ben Brubaker for helpful discussions.

References

[1] E. Artin, *Uber eine neue Art von L-Reihen*, Collected Papers, Springer, 1965.

[2] S. Gelbart, *An Elementary Introduction to the Langlands Program*, Bull. A.M.S. Vol. 10, No. 2 (1984), 177–220.

[3] W. Fulton and J. Harris, *Representation Theory*, Graduate Texts in Mathematics, Springer, 2004.

[4] M. R. Murty, *Selberg's Conjecture and Artin L-functions*, Bull. A.M.S. Vol. 31, No. 1 (1994), 1–14.

[5] J. Neukirch, *Algebraic Number Theory*, Grundlehren der mathematischen Wissenschaften, Springer, 1999.

[6] J.P. Serre, *Linear Representations of Finite Groups*, Graduate Texts in Mathematics, Springer, 1971.

[7] J. Tate, *Problem 9: the general reciprocity law*, Proc. Sympos. Pure Math., A.M.S., Vol. 28 (1976), 311–322.

STUDENT ARTICLE

2 Hilbert's Nullstellensatz and Schemes

Miles Dillon Edwards†
Indiana University '13
Bloomington, IN 47405
edwardmd@indiana.edu

Abstract

We prove Hilbert's Nullstellensatz from a geometric perspective, and use it to motivate the beginnings of scheme theory.

Disclaimer. In the rest of this paper, "ring" is shorthand for "commutative ring with unit".

2.1 Introduction, Statement, and Start of Proof

In differential geometry, we study manifolds—spaces that look like $\mathbb{R}^n$ in small neighborhoods. The simplest kinds are curves and surfaces. Level sets of smooth functions are also manifolds (possibly with singularities). So are the solution sets of systems of equations involving smooth functions.

In algebraic geometry, we study algebraic analogs. The classical objects of study are called *algebraic varieties*. These are solution sets to systems of algebraic equations, i.e., common zeroes of polynomials over some field K. Familiar examples include finite collections of lines, planes, hyperplanes, and even points, as well as conic sections, quadric surfaces, elliptic curves, and graphs of polynomials. We can consider solutions in different contexts. We may consider solutions in affine n-dimensional K-space, that is, K^n. However, it is often convenient to work in projective space, which allows certain statements that are "almost always" true (e.g., "any two distinct lines meet in one point") to become statements that are *always* true. Still, projective space is "locally" affine, in the sense that every point has a neighborhood (in fact, quite a large one—almost the entire space) that looks like affine space.

One pleasing result from classical algebraic geometry is Bézout's Theorem ([6], III.2), which says (in its elementary form) that any two curves of degrees m and n (in a projective plane over an algebraically closed field) meet in mn points, if multiplicity is counted carefully. This seems intuitive—after all, it matches the case where the curves are just collections of lines. However, it is not easy to prove, in part due to the fact that it is tricky to define multiplicity of intersection in a precise way. Another charming result is the fact that every cubic surface in projective space (over an algebraically closed field) contains 27 lines ([3], V.4).

Hilbert's Nullstellensatz is a result (or collection of results) that connects n-dimensional affine K-space (where K is an algebraically closed field) with its ring of (polynomial) functions in a precise and intimate way. Without further ado, we present the theorem—or at least, one version of the theorem.

Theorem 1 (Hilbert's Nullstellensatz, Version 1). *Let K be an algebraically closed field. Then the maximal ideals of the ring $K[x_1, \ldots, x_n]$ are precisely of the form $I(p)$, where $I(p)$ denotes the set of polynomials that vanish at the point p in n-dimensional K-space.*

Start of proof. We first note that $I(p)$ is always a maximal ideal of $K[x_1, \ldots, x_n]$, as it is in fact the kernel of the evaluation map $K[x_1, \ldots, x_n] \to K$ that evaluates polynomials at p. This map is clearly surjective, as $K[x_1, \ldots, x_n]$ contains all the constant polynomials. It follows that

†Miles Dillon Edwards is a sophomore studying mathematics and cello performance at Indiana University. He has taught at PROMYS and works for the Art of Problem Solving. He has particular interest for group theory, number theory, and most things algebraic.

$K[x_1, \ldots, x_n]/I(p)$ is a field isomorphic to K, so $I(p)$ is maximal. It thus remains to show that every maximal ideal of $K[x_1, \ldots, x_n]$ arises in this way.

This is the difficult part of the proof, but it is plausible. If we have a maximal ideal I of $K[x_1, \ldots, x_n]$, then the quotient field $K[x_1, \ldots, x_n]/I$ is canonically a field extension of K. This is suspicious, because K is supposed to be algebraically closed, and we are only working with finitely many variables. The only way things could possibly go wrong would be if we somehow got a transcendental extension of K. This seems highly unlikely, as such an extension would have to contain a transcendental t, as well as $1/P(t)$, for every nonzero polynomial P. Our proof involves a little more work, as we could have an extension that is not purely transcendental, but this is the main idea.

2.2 Rings of Integers and the End of the Proof

We now develop some commutative algebra that will be familiar to students of algebraic number theory.

Definition 2. Let A be a subring of a ring B. An element x of B is *integral over* A if it is a zero of a monic polynomial with coefficients in A; that is, if there are elements $a_0, \ldots, a_{n-1}$ in A such that

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0.$$

The motivation for this terminology comes from number theory, where B is usually a number field (i.e., a finite extension of the field of rationals) and A is usually $\mathbb{Z}$, the ring of ordinary integers. In this case, the algebraic integers are the zeroes of monic polynomials with integer coefficients.

It turns out that the elements of B that are algebraic over A constitute a ring, called the *integral closure of* A (in B). Furthermore, if B has no zero divisors, then every element of B that is algebraic over A is equal to some x/a, where x is an integer over A and a is an element of A. We now prove these facts.

Proposition 3 (Criterion for Integrality). *Let B be an extension of a ring A. An element x of B is integral over A if and only if $A[x]$ (a subring of B) is finitely generated as an A-module.*

Proof. First, suppose that $A[x]$ is a finitely generated A-module. Consider the ascending chain of submodules

$$A \subset A + Ax \subset A + Ax + Ax^2 \subset \cdots.$$

The union of this chain is the entire module $A[x]$. On the other hand, $A[x]$ is finitely generated. Thus if we fix a generating set, then there must be some finite n such that $A + Ax + \cdots + Ax^{n-1}$ contains each of the elements of the generating set of $A[x]$. But then this submodule $A + Ax + \cdots + Ax^{n-1}$ must be equal to $A[x]$—in particular, there must exist elements $a_0, \ldots, a_{n-1}$ of A such that

$$x^n = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1}.$$

It follows that x is integral over A.

Conversely, suppose that x is integral over A. Let

$$x^n + a_{n-1}x^{n-1} + \cdots + a_0$$

be a monic polynomial in x which evaluates to zero. Then from the polynomial division algorithm, every element of $A[x]$ is equal to some polynomial in x of degree at most $n-1$, with coefficients in A. Thus $A[x]$ is generated by the finite set $\{1, x^1, \ldots, x^{n-1}\}$. □

The next proposition resembles a familiar result from field theory, that dimension of field extensions is multiplicative.

Proposition 4. *Let B be a ring, and let A be a subring of B. Let C and D be subrings of B that contain A. If C and D are finitely generated as A-modules, then so is CD, the least subring of B containing C and D.*

Proof. Let $\{c_1, \ldots, c_m\}$ be a generating set for C, and let $\{d_1, \ldots, d_n\}$ be a generating set for D. We claim that the elements of the form $c_i d_j$ generate CD as an A-module. To this end, we note that the elements of CD are the elements of B that are sums of elements of the form cd, for c in C and d in D. It thus suffices to show that every element of the form cd lies in the submodule generated by the $c_i d_j$. For this, we note that for any c in C, there are elements $a_1, \ldots, a_m$ of A such that

$$c = \sum_{i=1}^{m} a_i c_i;$$

similarly, for any d in D, there are coefficients $a'_1, \ldots, a'_n$ such that

$$d = \sum_{j=1}^{n} a'_j d_j.$$

Then we have

$$cd = \sum_i a_i c_i \sum_j a'_j d_j = \sum_{i,j} (a_i a'_j) c_i d_j,$$

which lies in the submodule spanned by the $c_i d_j$. Thus the set of elements of the form $c_i d_j$ generate CD as an A-module; since the generating set in question is finite, we are done. □

Now we get the result we wanted from the previous two propositions.

Definition 5. A ring A is *Noetherian* if one of the following equivalent conditions is satisfied:

- Every ascending chain of ideals $I_0 \subset I_1 \subset \cdots$ is eventually constant.
- Every ideal of A is finitely generated.
- Every submodule of a finitely generated A-module is finitely generated.

Proposition 6 (The Set of Integral Elements is a Ring). *Let A be a subring of a ring B. Then the set of elements of B that are integral over A constitute a ring.*

Proof. Though this proposition is true generally, it simplifies the proof (and is sufficient for our purposes) to deal with the case where the ring A is *Noetherian*.

Suppose that x and y are integral over A. Then $A[x]$ and $A[y]$ are finitely generated A-modules. Since $A[x] = A[-x]$, it follows that $-x$ is integral over A. Now, $x + y$ and xy both belong to $A[x, y]$, which is the subring of B generated by $A[x]$ and $A[y]$. But the ring $A[x, y]$ is a finitely generated A-module, by proposition 4. Since A is Noetherian, the submodules $A[x + y]$ and $A[xy]$ are also finitely generated. Therefore $x + y$ and xy are integral over A. This gives us what we wanted. □

Proposition 7 (All Algebraic Elements Arise from Integral Elements). *Let B be a ring, and let A be a subring of B. Let y be an element of B that is algebraic over A. Then there is some nonzero element a of A and some element x of B, integral over A, such that $x = ay$.*

Proof. By definition, there are some elements $a_0, \ldots, a_n$ such that

$$\sum_{i=0}^{n} a_i y^i = 0,$$

with $a_n \neq 0$. Let $x = a_n y$. Then we have

$$0 = a_n^{n-1} \sum_{i=0}^{n} a_i y^i = (a_n y)^n + \sum_{i=0}^{n-1} a_i (a_n)^{n-1-i} (a_n y)^i = x^n + \sum_{i=0}^{n-1} (a_i a_n^{n-1-i}) x^i.$$

Then x is integral over A, and $x = a_n y$, where a_n belongs to A, as desired. □

We finish with a result that allows us to determine concretely whether elements of B are integral over A, if B is a field.

Proposition 8 (Concrete Criterion for Integrality). *Let B be a field, and let A be a subring of B which is a unique factorization domain. Let C be the field of fractions of A. Then an element x of B is integral over A if and only if it is algebraic over C, and all the coefficients of its (monic) minimal polynomial (over C) lie in A.*

Proof. It is clear that if the latter condition holds, then x is integral. For the converse, suppose that x is integral; then by definition, x is algebraic over C. Let f be the (monic) minimal polynomial of x, and let h be a monic polynomial with coefficients in A such that $h(x) = 0$. Then there must be some g with coefficients in C such that $fg = h$. But by Gauss's lemma, both f and g must have coefficients in A. In particular, f, the minimal polynomial of x, has coefficients in A. □

Conclusion of the Proof of the Nullstellensatz. Suppose for the sake of contradiction that I is an ideal of $K[x_1, \ldots, x_n]$ such that $K[x_1, \ldots, x_n]/I$ is a transcendental extension L of K. We pick a transcendence base $t_1, \ldots, t_k$ for L, i.e., a maximal set of elements with no polynomial relations among them. (We can even take the t_i to be a subset of the x_i. For more details on transcendence bases, we refer the reader to section VIII.1 of [4].) Then L is an algebraic (in fact, finite) extension of the purely transcendental extension $K(t_1, \ldots, t_k)$. Let A be the subring $K[t_1, \ldots, t_k]$. Then A is a unique factorization domain. (The interested reader may find a proof of this fact in Theorem 2.3 of [4], in chapter IV, related to Gauss's lemma.) It is also a Noetherian ring, by Hilbert's Basis Theorem ([1], 7.5; [4], section IV.4). Since all algebraic elements in this situation arise from integral elements, every element of L is of the form x/a, where x is integral over A and a is an element of A, i.e., a polynomial in $t_1, \ldots, t_k$ with coefficients in K.

Now, let $a_1, \ldots, a_n$ be nonzero elements of A such that the elements $a_i x_i$ (of L) are all integral over A. We claim that the a_i cannot all be units (i.e., elements of K)—indeed, otherwise, the x_i (or rather, their images in L) would all be algebraic integers over A, so all the elements of L would be integral over A. This implies in particular that $1/t_1$ is not in L, as its minimal polynomial is $P(z) = z - 1/t_1$, in violation of our concrete criterion for integrality. Thus the a_i are not all units—that is, they are not all constant polynomials. Then their product $a_1 \cdots a_n$ is not constant, so $a_1 \cdots a_n + 1$ is not zero. Consider, then, the element

$$\frac{1}{a_1 \cdots a_n + 1}$$

of L. Since L is a quotient of $K[x_1, \ldots, x_n]$, this element must be equal to some polynomial in $x_1, \ldots, x_n$. It follows that there must be non-negative exponents $e_1, \ldots, e_n$ such that

$$\frac{a_1^{e_1} \cdots a_n^{e_n}}{a_1 \cdots a_n + 1}$$

is integral over $A = K[t_1, \ldots, t_k]$. But this cannot be—indeed, this element belongs to the field of fractions of A, but not to A (since the numerator and denominator have no common factors, and the denominator is not a unit). Therefore it cannot be integral over A, from our concrete criterion. This is a contradiction, so L must be an algebraic extension. □

If we examine our methods carefully, we see that we only used the algebraic closure of K when we noted that all algebraic extensions of K are trivial. Thus the following, slightly more general version of the theorem is true.

Theorem 9 (Hilbert's Nullstellensatz for a General Field). *Let K be a field, and let I be a maximal ideal of the polynomial ring $K[x_1, \ldots, x_n]$. Then the quotient $K[x_1, \ldots, x_n]/I$ is a finite extension of K.*

2.3 The Strong Form

The Nullstellensatz we have given is called the "weak version". The "strong version" is as follows.

Theorem 10 (Hilbert's Nullstellensatz, Strong Version). *Let K be an algebraically closed field, and let $\mathfrak{a}$ be an ideal of $K[x_1, \ldots, x_n]$. Let $V(\mathfrak{a})$ be the variety in K^n of points at which all the elements of $\mathfrak{a}$ vanish. Then the ideal of elements that vanish on $V(\mathfrak{a})$ is the radical of $\mathfrak{a}$.*

We recall that the radical of an ideal $\mathfrak{a}$ is the collection of elements a such that a^n belongs to $\mathfrak{a}$ for some positive n; it is an ideal, and it is sometimes denoted $r(\mathfrak{a})$, or $\sqrt{\mathfrak{a}}$. If we use $I(V)$ to denote the ideal of elements vanishing on a variety V, then we can express the strong form as a pithy equation :

$$I(V(\mathfrak{a})) = \sqrt{\mathfrak{a}}.$$

In our proof, we use the fact that the radical of an ideal is the intersection of the prime ideals containing that ideal. This is equivalent to the statement that the nilradical of a ring (i.e., the collection of nilpotent elements) is the intersection of the ring's prime ideals. We prove this fact in our appendix; the reader may also refer to [1], chapter 1.

Proof of the Strong Version. It follows from our definitions that if f^n belongs to $\mathfrak{a}$, then f must vanish at every point in $V(\mathfrak{a})$. Thus the radical of $\mathfrak{a}$ is a subset of the ideal of $V(\mathfrak{a})$, so it suffices to show that the opposite inclusion holds. This reduces to showing that if f is a polynomial that vanishes on $K[x_1, \ldots, x_n]$, then f belongs to every prime ideal containing $\mathfrak{a}$, from the characterization of the nilradical of the ring $K[x_1, \ldots, x_n]/\mathfrak{a}$.

To this end, let $\mathfrak{p}$ be a prime ideal containing $\mathfrak{a}$, and let $\mathfrak{q}$ be the ideal generated by $\mathfrak{p}$ in the ring $K[x_1, \ldots, x_n, y]$. We note that the variety of the ideal $(\mathfrak{q}, 1 - yf)$ has no points (in K^{n+1}); indeed, if a point vanishes under all the elements of $\mathfrak{q}$, then it must vanish under f, so $1 - yf$ must evaluate to 1. By the weak version of the Nullstellensatz, this means that there is no maximal ideal that contains $(\mathfrak{q}, 1 - yf)$, so $K[x_1, \ldots, x_n, y]/(\mathfrak{q}, 1 - yf)$ must be the zero ring. But this ring is just the localization of the ring $K[x_1, \ldots, x_n]/\mathfrak{p}$ by the element f. Since this quotient ring has no zero divisors (as $\mathfrak{p}$ is a prime ideal), this means that f must be zero in this ring—that is, $\mathfrak{p}$ must contain f. □

This theorem shares its name with the "weak" form because the weak form follows from the strong form thus: if an ideal $\mathfrak{a}$ has no points on its variety, then its radical must be the collection of polynomials that vanish on the empty set, which is (vacuously) the collection of all polynomials in our ring. In particular, $\mathfrak{a}$ cannot be a maximal (proper) ideal, as its radical is the unit ideal; it follows that the only maximal ideals are those that arise from points.

As we have seen, the other direction is not as easy. The idea of introducing a variable y to stand for $1/f$ is called the *Rabinowitsch trick*, after J. L. Rabinowitsch, who published it in a one-page paper in 1929. There are many ways of proving the weak form of the Nullstellensatz, but modern proofs of the strong form seem, as a rule, to deduce the strong form from the weak form using this trick. Curiously, little seems to be known about Rabinowitsch—he is not currently listed in the Mathematics Genealogy Project. Rabinowitsch's paper was apparently submitted from Moscow, but this author has not succeeded in finding any other information about him.

The method is called a "trick", but it seems a little less arbitrary if we are aware of the Jacobson radical of a ring R, which is both the intersection of all maximal ideals of R and the collection of elements x such that $1 - yx$ is invertible for every element y of x ([1], ch. 1). We know that f belongs to the Jacobson radical of $K[x_1, \ldots, x_n]/\mathfrak{a}$, and we want to wreak havoc on any prime ideals $\mathfrak{p}$ to which f does not belong. And what could cause greater havoc than requiring $1 - yf$ to be zero, when it is supposed to be invertible?

2.4 Historical Aside

The German word *Nullstellensatz* roughly means "theorem of the zeroes". The reason for this name becomes more apparent if we state the classical versions of the theorems. These are equivalent to the versions we have seen so far; they have the advantage of making the theorem appear more striking and making the connection between the two versions more apparent, but the disadvantage of obscuring the geometric intuition.

Theorem 11 (Hilbert's Nullstellensatz, Classical Weak Form). *Let K be an algebraically closed field, and let $f_1, \ldots, f_r$ be polynomials in n variables over K. Then either the f_i have a common zero, or there exist polynomials $g_1, \ldots, g_r$ such that*

$$\sum_{1 \leq i \leq r} g_i f_i = 1.$$

Theorem 12 (Hilbert's Nullstellensatz, Classical Strong Form). *Let K be an algebraically closed field,and let $f_1, \ldots, f_r$ be polynomials in n variables over K. Let f be another such polynomial. If f vanishes on all the common zeroes of the f_i, then there exist polynomials $g_1, \ldots, g_r$ and an integer m such that*

$$\sum_{1 \leq i \leq r} g_i f_i = f^m.$$

When we express the theorems like this, the weak form is the special case of strong form where f is the constant polynomial 1.

2.5 Varieties and Schemes

Hilbert's Nullstellensatz gives us some hope of expressing results from algebraic geometry more purely in the language of commutative algebra—points correspond to maximal ideals, and more generally, sub-varieties correspond to (radical) ideals. We can think of a ring of the form

$$K[x_1, \ldots, x_n]/\mathfrak{a}$$

as the ring of functions on the variety $V(\mathfrak{a})$. Maximal ideals of this ring correspond to maximal ideals of $K[x_1, \ldots, x_n]$ that contain the ideal $\mathfrak{a}$—i.e., points on the variety of $\mathfrak{a}$. In general, ideals correspond to sub-varieties.

But why limit ourselves to finitely generated rings over algebraically closed fields? It is sometimes useful to work over fields that are not algebraically closed—indeed, in number theory, we often want to work over $\mathbb{Q}$ or finite extensions of $\mathbb{Q}$. So we could be tempted to define an abstract variety as the collection of maximal ideals of an arbitrary (commutative) ring. The problem with this idea is that ring homomorphisms do not induce convenient general relations between maximal ideals of rings. But they do induce nice relations between *prime* ideals: specifically, the inverse images of prime ideals under ring homomorphisms are again prime ideals. This motivates the definition of the *spectrum of a ring* as the set of all prime ideals of the ring; it comes with a topology, where a closed set is the collection of prime ideals containing a given ideal. In the case where our ring is $K[x_1, \ldots, x_n]$, these are just the varieties of n-space; intuitively, closed sets of the spectrum are (sub)varieties of the space.

This gives us a generalized notion of affine varieties. For the generalizations of projective space and other things, we allow spaces in which every point has a neighborhood isomorphic to some affine space, i.e., to the spectrum of some ring, in such a way that the isomorphisms. This construction is done precisely, and in more detail in [3], II.1–2.

This definition is slightly strange, in part because of the presence of non-maximal prime ideals. These are "points" in the spectrum, but they also define closed sets containing other points. As it turns out, the closed sets they define are irreducible; that is, they cannot be expressed as a union of two strictly smaller closed sets. We think of these ideals as generic points on these irreducible closed subsets. It turns out that this relation is bijective; every irreducible closed subset has a generic point that is a prime ideal (see [1], ex. 19; [3], III, 3.1).

Another strange aspect of this definition is that our space is endowed with a very weak topology. We have already noted that there may be many "generic" points whose closures contain other points. Our open sets will be very large—after all, in our model case, affine n-space, our closed sets have dimension smaller than the space itself. For example, the spectrum of $\mathbb{C}[x]$ consists of the points of the complex plane (corresponding to maximal ideals), along with a generic point of the entire space, corresponding to the zero ideal. But the only open sets are those obtained by removing

finitely many closed points. This means that, among other things, every injective function from any Hausdorff space into the spectrum of $\mathbb{C}[x]$ is continuous, as long as we avoid the generic point! This strange topology makes standard topological tools difficult to use, but certain cohomology theories have been developed for use with schemes and related spaces; some of these theories are developed in [3], chapter III. This theory of schemes seems strange at first, but it has been influential, and it lies behind some of the more spectacular recent advances in number theory.

2.6 Acknowledgements

I am indebted to many teachers and friends who have helped me lately. I owe particular debts to Michael Larsen of Indiana University, and to my parents, Meg Dillon and Steve Edwards, of Southern Polytechnic State University.

2.7 Appendix: the Nilradical

Proposition 13. *Let R be a ring. Then the nilradical of R is the intersection of the prime ideals of R.*

Proof. Suppose first that an element a belongs to the nilradical of R. Then $a^n = 0$ for some positive integer n; it follows that a must be zero in any quotient of R that has no zero divisors. Thus a lies in every prime ideal of R.

Suppose on the other hand that $a^n \neq 0$ for every positive integer n. Consider the set S of ideals that avoid all powers of a, ordered by inclusion. This set is non-empty (since it contains the zero ideal), and the union of any totally ordered family of elements of S also belongs to S. Thus S satisfies the hypotheses of Zorn's lemma, so it has a maximal element; let $\mathfrak{p}$ be such a maximal element. We claim that $\mathfrak{p}$ is a prime ideal.

Indeed, suppose that x and y are elements of R such that $xy = 0$ in $R/\mathfrak{p}$. Since $\mathfrak{p}$ is a maximal element of S, we know that either $\mathfrak{p}$ contains x, or the ideal $\mathfrak{p} + (x)$ contains a^n for some integer n. Similarly, either $\mathfrak{p}$ contains y, or the ideal $\mathfrak{p} + (y)$ contains a^m for some integer m.

Suppose that powers of a belong both to $\mathfrak{p} + (x)$ and to $\mathfrak{p} + (y)$. Then there exist elements d and e of R such that $dx \equiv a^n \pmod{p}$ and $ey \equiv a^m \pmod{p}$. Then we have

$$0 \equiv dxye \equiv a^{m+n} \pmod{p},$$

a contradiction. Therefore one of x and y must belong to $\mathfrak{p}$. Thus $\mathfrak{p}$ is a prime ideal that avoids all powers of a.

Thus if a does not belong to the nilradical of R, then R has a prime ideal not containing a, so we are done. □

References

[1] M. F. Atiyah and I.G. MacDonald: *Introduction to Commutative Algebra*. Westview Press, 1969.

[2] D. Eisenbud: *Commutative Algebra with a View Toward Algebraic Geometry*. New York: Springer, 1999.

[3] R. Hartshorne: *Algebraic Geometry*. New York: Springer, 1977.

[4] S. Lang: *Algebra*. New York: Springer, 2002.

[5] J. L. Rabinowitsch: *Zum Hilbertschen Nullstellensatz*, Math. Ann. **102** (1930) #1, 520.

[6] I. R. Shafarevich (trans. K. A. Hirsch): *Basic Algebraic Geometry*. Springer, 1977.

STUDENT ARTICLE

3 Toronto Spaces

Manuel Rivera†
Massachussets Institute of Technology '10
Cambridge, MA 02138
manuelr@mit.edu

Abstract

A topological space X is said to be a Toronto space if it is homeomorphic to all its subspaces of the same cardinality. In this paper, we present some results in the area of set theoretic topology concerning these spaces and the so-called Toronto problem, an open question about the existence of a Toronto space that is uncountable, non-discrete and Hausdorff. We conclude the paper that such a space, if it exists, must be separable (assuming the Continuum Hypothesis).

3.1 Introduction

One often hears in mathematical circles that the field of point set topology is dead. It is certainly not a popular area of research; however, there are many interesting open problems which are accessible to young researchers and have connections with other areas of mathematics, such as logic and set theory. In this paper, we present some results and questions about Toronto spaces, assuming only the basic notions of point-set topology. We begin by reviewing some preliminary concepts before describing the nature of Toronto spaces.

Definition 1. If X is any set, the collection of all subsets of X is a topology on X which we call the discrete topology. We say a topological space X is non-discrete if there exists a subset of X which is not an open set in X.

Definition 2. Let X be a topological space with topology τ. If Y is a subset of X, the collection $\tau' = \{Y \cap O : O \in \tau\}$ is a topology on Y, called the subspace topology.

Now, we define our main object of study, which was introduced by J. Steprans in [4]:

Definition 3. Let X be a topological space. We say X is a Toronto space if it is homeomorphic to all its subspaces of the same cardinality as X.

In other words, we say a topological space X is a Toronto space if for any subspace $Y \subset X$ such that $|Y| = |X|$, Y is homeomorphic to X ($Y \cong X$).

Note that any set X with the discrete topology is a Toronto space. For if $Y \subset X$ then the subspace and the discrete topologies on Y are equivalent. Hence, if $|Y| = |X|$, any bijection between X and Y induces a homeomorphism. However, not every Toronto space must have the discrete topology. For example, let Z be any set and consider the collection $\tau = \{A \subset Z : A = \emptyset$ or $Z - A$ is finite$\}$. It is straightforward to check that τ is a topology on Z; we call it the finite complement topology. A space with the finite complement topology is a Toronto space, since if X has the finite complement topology so does any subspace.

We can narrow our study by considering Hausdorff Toronto spaces. First, recall the definition of a Hausdorff space:

†Manuel Rivera graduated from MIT in 2010 with an undergraduate degree in mathematics. He is currently a graduate student at the Graduate Center of The City University of New York working in algebraic and geometric topology. In his free time, he enjoys thinking about problems in mathematical logic, set theory and combinatorics.

Definition 4. A topological space X is Hausdorff if for any two distinct points x and y of X, there exist open sets U_x and U_y in X such that $x \in U_x$, $y \in U_y$, and $U_x \cap U_y = \emptyset$

In the finite complement topology, each open set contains all but finitely many points, so if the space is infinite, any two open sets intersect. Hence, an infinite topological space with the finite complement topology is not Hausdorff. It is much harder to find examples of non-discrete Hausdorff Toronto spaces. We will show that every countable Hausdorff Toronto space is discrete. For the uncountable case, the question is still open. This is known as the *Toronto problem*:

Question 1. Is there an uncountable, non-discrete, Hausdorff Toronto space?

3.2 Countably Infinite Toronto Spaces

As usual, we denote by ω the least infinite ordinal, by ω_1 the first uncountable ordinal and by $\aleph_0$ and $\aleph_1$ their respective cardinalities. From now on, we are interested in Hausdorff Toronto spaces. In this section we will prove that the set of isolated points of any Hausdorff Toronto space of infinite cardinality is infinite. From this, it will follow that any Hausdorff Toronto space of cardinality $\aleph_0$ is discrete. Recall the following definition and notation:

Definition 5. Let X be a topological space. We say a point $x \in X$ is isolated if the set $\{x\}$ is open in X.

If X is a topological space, we will denote by X^* the set of all isolated points in X. Now we can prove our first theorem: the existence of an infinite set of of isolated points in any infinite Hausdorff Toronto space.

Theorem 6. *If X is an infinite Hausdorff Toronto space, then the set X^* is infinite.*

Proof. First, let us show that X^* is non empty. Let x and y be two distinct points in X. Since X is Hausdorff, there exist disjoint open sets U and V such that $x \in U$ and $y \in V$. Let λ denote the cardinality of X, i.e. $\lambda = |X|$. Consider the set $A = X - U$. It follows that we have either $|A| < \lambda$ or $|A| = \lambda$.

If $|A| < \lambda$, then we know by cardinal arithmetic that $|U \cup \{y\}| = \lambda$. Consider the subspace $Y = U \cup \{y\}$ of X, and note that $V \cap Y = \{y\}$. Hence, $\{y\}$ is an open set in Y. We know $|Y| = \lambda = |X|$ and that X is a Toronto space, so there exists a homeomorphism $h : Y \to X$. Since $\{y\}$ is open in Y it follows that $\{h(y)\}$ is open in X. Therefore, the point $h(y)$ is isolated in X, so $h(y) \in X^*$.

Now, suppose that $|A| = \lambda$. Consider the subspace $Z = A \cup \{x\}$ of X. It follows, again by cardinal arithmetic, that $|Z| = \lambda$, so we can apply an argument similar to the previous case. Note that $U \cap Z = \{x\}$, so the set $\{x\}$ is open in Z. Since $|Z| = \lambda = |X|$ and X is a Toronto space we have that there is a homeomorphism $h : Z \to X$. It follows that $\{h(x)\}$ is open in X, so $h(x) \in X^*$.

We have proved that X^* is nonempty; now we show that it contains an infinite number of points. For the sake of a contradiction, suppose X^* is finite. Let $X^* = \{x_1, ..., x_k\}$. Consider the subspace $W = X - \{x_1\}$ of X. By cardinal arithmetic we know that $|W| = \lambda = |X|$. Hence, since X is a Toronto space there is a homeomorphism $h : X \to W$. It now follows that the only isolated points in W are $h(x_1), ..., h(x_k)$, so $W^* = \{h(x_1), ..., h(x_k)\}$. Also, $\{x_i\} \cap W = \{x_i\}$ for $i = 2, ..., k$ and since $\{x_i\}$ is open in X for all $i = 1, ..., k$, we have that the points $x_2, ..., x_k$ are isolated in W. We also know that W has exactly k isolated points, so let x_0 be the remaining isolated point of W which is not in the set $\{x_2, ..., x_k\}$. By the definition of the subspace topology, there is an open set O in X such that $O \cap W = \{x_0\}$. Hence, $O - \{x_0\} \subset X - W = \{x_1\}$. So, we have two cases: either $O - \{x_0\} = \emptyset$ or $O - \{x_0\} = \{x_1\}$.

If $O - \{x_0\} = \emptyset$ then the point x_0 would be an isolated point in X not in the set $X^* = \{x_1, ..., x_k\}$, contradicting the fact that X^* is the set of all isolated points in X. If $O - \{x_0\} = \{x_1\}$ then $O = \{x_0, x_1\}$. Since X is Hausdorff, the set $\{x_1\}$ is closed in X. But, $\{x_0\} = O \cap (X - \{x_1\})$, and since both O and $X - \{x_1\}$ are open in X it follows that $\{x_0\}$ is open in X as well, thus $x_0 \in X^*$. This contradicts with the fact that $X^* = \{x_1, ..., x_k\}$.

Both cases have led to a contradiction, so it follows that the set X^* contains an infinite number of points. □

Now, we can completely classify countably infinite Hausdorff Toronto spaces. As a direct consequence of Theorem 6, we have that these spaces must have the discrete topology.

Theorem 7. *If X is a Hausdorff Toronto space of cardinality $\aleph_0$, then X has the discrete topology.*

Proof. Let X be a countably infinite Hausdorff Toronto space. By Theorem 6, we have that X^* is an infinite subset of X. Therefore, $|X| = \aleph_0 = |X^*|$, and since X is a Toronto space, it follows that $X \cong X^*$. Thus, every point in X is isolated, so X has the discrete topology. □

3.3 Uncountable Toronto Spaces

We have pinned down, up to homeomorphism, countably infinite Hausdorff Toronto spaces. The natural generalization is to consider higher cardinalities. We will restrict out attention to spaces of cardinality $\aleph_1$. There has not been much research on Toronto spaces of cardinality greater than $\aleph_1$, since this case already poses a difficult problem, i.e. Question 1. As the reader may suspect, when studying spaces of cardinality $\aleph_1$, we need to assume the Continuum Hypothesis to obtain results. Recall the Continuum Hypothesis (CH): There is no set S satisfying $\aleph_0 < |S| < 2^{\aleph_0}$, or equivalently $2^{\aleph_0} = \aleph_1$.

Theorem 8. *If X is a non-discrete Hausdorff Toronto space of cardinality $\aleph_1$, then, assuming CH, X is separable.*

Proof. Let us prove that X^* is a countable dense subset of X. Denote by $\overline{X^*}$ the closure of X^*, so we must show $\overline{X^*} = X$.

We claim that $|\overline{X^*}| = \aleph_1$. By Theorem 7, we have that $\overline{X^*}$ is not finite and if $|\overline{X^*}| < \aleph_1$, then, assuming CH, $|\overline{X^*}| = \aleph_0$. But then, $|X - \overline{X^*}| = \aleph_1 = |X|$ implying that X is homeomorphic to $X - \overline{X^*}$ which is a contradiction, since X contains isolated points and $X - \overline{X^*}$ does not.

Since $|\overline{X^*}| = \aleph_1 = |X|$, there is a homeomorphism $h : \overline{X^*} \to X$. Note that $X = h(\overline{X^*}) = \overline{h(X^*)} \subseteq \overline{X^*}$, as desired. Thus, X^* is a dense subset of X.

Moreover, we can use a similar argument to show X^* is countable. If not, then by CH, $|X^*| = \aleph_1 = |X|$ and therefore X is homeomorphic to X^*. Thus, since all points in X^* are isolated, it follows that X must have the discrete topology, contradicting the initial hypothesis. □

Note that in the proof above we only used the fact that X is non-discrete when proving that X^* is countable. So, the set of isolated points of a Hausdorff Toronto space X of cardinality $\aleph_1$ is a dense set. If we require X to be non-discrete, then X^* is countable.

3.4 Conclusion

Theorem 8 is the only known result about Hausdorff Toronto spaces of cardinality $\aleph_1$. It appears that more advanced tools from set theory are required to study deeper properties of Hausdorff topologies on ω_1. There are some results about Hausdorff spaces of cardinality $\aleph_1$, assuming properties weaker than the Toronto property. The deepest result, as far as we know, is due to Soukup [3] who proved that it is consistent that there is a hereditarily separable, 0-dimensional, Hausdorff space X of cardinality $\aleph_1$ such that for each uncountable subspace Y of X there is a continuous bijection $\phi : Y \to X$ and there is a partition $(Y_i)_{i<n}$ of Y into finitely many pieces such that ϕ is a homeomorphism when restricted to Y_i for each $i < n$. The proof is dense and it relies mainly on set theoretic tools. However, it looks like Question 1 might be studied using more topological tools. This short note is intended to spark interest in the Toronto Problem among young researchers, and to prevent the area of set theoretic topology from being forgotten with time.

References

[1] N. Hernández. *Remarks about Toronto spaces.* Divulg. Mat., (1998) 6(2), 87-91.

[2] J. Munkres. *Topology: A First Course.* N.J.: Prentice-Hall Inc., Englewood Cliffs, 1975.

[3] L. Soukup. *A piecewise Toronto space.* Studia Sci. Math. Hungar. (2004) 41(3), 325-337, .

[4] J. Steprāns. Steprāns' problems. In *Open Problems in Topology.* North-Holland, Amsterdam, 1990.

STUDENT ARTICLE

4

An Introduction to Sieve Theory

Seth Neel†
The Wheeler School '11
Providence, RI 02906
twinlsn@yahoo.com

Abstract

We introduce the two simplest combinatorial sieves: the ancient Sieve of Eratosthenes and the related Sieve of Legendre. We then use the ideas developed to explore the Brun Sieve, proving Brun's bound on the number of twin primes less than N. Finally we describe the more modern Selberg Sieve, and highlight some of the major contributions of sieve theory.

4.1 Introduction

Sieve theory is a set of tools in number theory that estimate (bound) the size of sets of *sifted integers*. Sifted integer sets are sets that do not follow a known pattern. For example, the set of primes less than N is the most basic example of a sifted set. Our concept of an arithmetical sieve is based on the following principle: if we take an integer sequence A, a set of primes B, and a number $z \geq 2$, and sift out from A all numbers divisible by primes $p \in B, p \leq z$, then the remaining integers in A can only have prime divisors from B greater than z. The object of Sieve theory is to estimate $S(A, B, z)$, the number of *unsifted* elements of A. We first examine the most basic sieve, the Sieve of Eratosthenes, which is used to generate the prime numbers. We then examine the related Legendre Sieve which uses the principle of inclusion-exclusion (PIE) to obtain a precise count of $S(A, P, z)$. We move on to the more complicated Brun Sieve which uses PIE and a simple inequality to bound $S(A, P, z)$, and use this bound to prove a bound on the number of twin primes less than N. Finally we discuss the Selberg Sieve and the state of modern sieve theory.

4.2 The Sieve of Eratosthenes and the Legendre Sieve

The primes are the simplest example of a sifted set, and accordingly the simplest example of a sieve is the Sieve of Eratosthenes; an ancient method for generating all the primes up to an integer N. The algorithm is summarized below:

1. List all the consecutive integers from two to N: $2, 3, \ldots, N$.
2. Let $p = 2$ be the first prime number, and strike from the list all multiples of p greater than p.
3. Find the first number on the list greater than p; this is the next prime; let p equal this number.
4. Repeat steps 2 and 3 until $p^2 > N$; all remaining unsifted numbers are prime.

Note that in the Sieve of Eratosthenes A is the set of naturals up to N, B is the set of primes p such that $p \leq \lfloor \sqrt{N} \rfloor$, and $z = \lfloor \sqrt{N} \rfloor$. It is easy to see that any unsifted numbers are prime because if they are not prime they must have two factors greater than $\lfloor \sqrt{N} \rfloor$ which means they are greater than N and thus not in the set. Now let us try to count the primes in the interval $[1, N]$

†Seth Neel is a senior at The Wheeler School in Providence, Rhode Island. He has taken math courses at Brown University since his junior year, and in the summer of 2009 he attended the Program in Mathematics for Young Scientists at Boston University. His interests include squash, chess, and rap music.

explicitly, which will take us from the Sieve of Eratosthenes to the related Sieve of Eratosthenes-Legendre or the Legendre Sieve. While the set of primes does not have enough structure to be easily counted, there are other sets in $[N, 2N]$ that we can count with relative ease; for example $|\{n \in [N, 2N] : n \equiv a \pmod{q}\}| = \frac{n}{q} + R_q$. Here the error term R_q depends on what N is modulo q, but it is quite small as long as q is small relative to N; when q gets large we can even have $R_q > \frac{n}{q}$.

Theorem 1. *The principle of inclusion-exclusion states that given finite sets* $B_1, B_2, \ldots, B_n$

$$|B_1 \cup B_2 \cup \cdots \cup B_n| = \left(\sum_{i=1}^{n} |B_i|\right) - \left(\sum_{1<i<j\leq n} |B_i \cap B_j|\right) + \cdots + (-1)^{n-1}|B_1 \cap \cdots \cap B_n|.$$

Example 2. Let us count all integers $k \in \mathbb{N}, k \leq n$ coprime to 3, 5. Because 3 and 5 are primes, k is coprime to them both if $3, 5 \nmid k$. The number of integers less than n divisible by p is $\lfloor \frac{n}{p} \rfloor$. Thus by inclusion-exclusion $|\{k \in \mathbb{N} | k \leq n, k \text{ coprime to } 3, 5\}| = n - \lfloor \frac{n}{3} \rfloor - \lfloor \frac{n}{5} \rfloor + \lfloor \frac{n}{15} \rfloor$.

The Sieve of Legendre simply counts the primes in [1,N] by sifting out the numbers divisible by primes using inclusion-exclusion. More formally if we have $A = \{n : n \leq N\}$, and $B = \{p : p \leq z\}$, the Legendre Sieve states:

$$S(A, B, z) = N - \sum_{p\leq z} \left\lfloor \frac{N}{p} \right\rfloor + \sum_{p_1<p_2\leq z} \left\lfloor \frac{N}{p_1 p_2} \right\rfloor - \sum_{p_1<p_2<p_3\leq z} \left\lfloor \frac{N}{p_1 p_2 p_3} \right\rfloor + \cdots$$

where the expansion has $2^{\pi(z)}$ terms, where $\pi(z)$ is the prime counting function. If we use $\frac{x}{q}$ to approximate the term $\lfloor \frac{x}{q} \rfloor$ we can have very large error in the Legendre Sieve due to the large number of terms in the sieve and the fast growth rate of the product of the primes less than N. However, we can still obtain some simple bounds for $S(A, B, z)$. The obvious lower bound (known as the *union bound*) is

$$S(A, B, z) \geq N - \sum_{p\leq z} \left\lfloor \frac{N}{p} \right\rfloor$$

and the slightly less obvious upper bound is

$$S(A, B, z) \leq N - \sum_{p\leq z} \left\lfloor \frac{N}{p} \right\rfloor + \sum_{p_1<p_2\leq z} \left\lfloor \frac{N}{p_1 p_2} \right\rfloor.$$

In general from the notion of inclusion-exclusion we can see that if we take the first n terms in the Legendre Sieve (excluding N) we get a lower bound for n even and an upper bound for n odd.

Definition 3. The Mobius function is defined by $\mu(d) = 0$ if $\exists p$ a prime such that $p^2 | d$, and $\mu(d) = -1^u$, where u is the number of prime divisors of d if not.

Then if we take $P = \prod_{p\leq z} p$, we can rewrite the Legendre sieve more concisely as:

$$S(A, B, z) = \sum_{d|p} \mu(d) \left\lfloor \frac{N}{d} \right\rfloor$$

4.3 The Brun Sieve and Brun's Theorem

We give part of Brun's proof that the sum of the reciprocals of the twin primes converges. A twin prime is a prime p such that $p + 2$ is prime (or alternatively $p - 2$). We will prove an upper bound for the number of twin primes less than N, which is an important step in Brun's proof. In the proof we prove an upper bound $S(A, P, z)$ or the number of elements in A with no prime factors greater than z which is of course a much larger set than the set of all twin primes in A.

Theorem 4 (Brun). *Let $\pi_2(N)$ denote the number of twin primes less than N. Then*

$$\pi_2(N) \ll \frac{N(\log\log N)^2}{(\log N)^2}$$

Proof. We make a simple observation: if $p, p+2$ are twin primes than $p(p+2)$ has no small prime factors; i.e. no prime factors less than p. Now let $f(x) = x(x+2)$ and our sequence $A = f(1), f(2), \ldots, f(N)$. Let P be the set of primes $p_1 < p_2 < p_3 < \cdots$ and A_d denote the elements of A that are divisible by d. Then inclusion-exclusion (the Legendre Sieve) gives us:

$$S(A,P,z) = \sum_{s \in \mathbb{N} \cup 0} (-1)^s \sum_{1<i_1<\cdots<i_s} A_{p_{i_1}p_{i_2}\cdots p_{i_s}} \tag{4.1}$$

As we saw earlier A_d is difficult to count when d is large relative to z; the error term becomes significant. When $d = p$ we have $A_p = \frac{e(p)N}{p} + R_p$ where $e(p) = 2$ for $p > 2$. This is because we will have roughly two times as many terms that are divisible by p in A than if we were sifting $A' = 1, 2, \ldots, N$ because numbers d congruent to 0 or -2 modulo p will have $f(d) \equiv 0 \pmod{p}$ and $e(p) = 1$ for $p = 2$. $R_p \leq 2$ is the error term. We can extend this for $d = \prod p_{i_1}p_{i_2}\cdots p_{i_s}$ to get

$$A_d = \frac{e(p_{i_1})e(p_{i_2})\cdots e(p_{i_s})}{d} N + R_d, R_d \leq 2^s \tag{4.2}$$

But now any attempt to substitute into (4.1) is futile because of the large error term R_d if s is at all large. The crux of Brun's way to get around this problem relies on the fact that we can truncate (4.1) at an even integer t, and we will get an upper bound for $S(A,P,z)$. Substituting into (4.1) we get

$$S(A,P,z) \leq \sum_{s=0}^{t} (-1)^s \sum_{1<i_1<\cdots<i_s} A_{p_{i_1}p_{i_2}\cdots p_{i_s}} \tag{4.3}$$

Substituting (4.2) into (4.3) we get

$$S(A,P,z) \leq N \sum_{s=0}^{t} (-1)^s \sum_{1<i_1<\cdots<i_s} \frac{e(p_{i_1})e(p_{i_2})\cdots e(p_{i_s})}{p_{i_1}p_{i_2}\cdots p_{i_s}} + \sum_{s=0}^{t} \binom{t}{s} 2^s \tag{4.4}$$

We use heuristics to point us in the right direction. Let us guess at the value of $S(A,P,z)$. We assume that $p|n \in A$ with probability $\frac{e(p)}{p}$, and that these events are independent over different p (we are assuming that the error R_p in the expression for A_p is 0). Then we have $S(A,P,z) = N\prod_{p<z}\left(1 - \frac{e(p)}{p}\right)$. So we perform manipulations to make $N\prod_{p<z}\left(1 - \frac{e(p)}{p}\right)$ the main term of $N\sum_{s=0}^{t}(-1)^s \sum_{1<i_1<\cdots<i_s} \frac{e(p_{i_1})e(p_{i_2})\cdots e(p_{i_s})}{p_{i_1}p_{i_2}\cdots p_{i_s}} + \sum_{s=0}^{t}\binom{t}{s}2^s$:

$$\prod_{p<z}\left(1 - \frac{e(p)}{p}\right) = \sum_{s=0}^{\pi(t)} (-1)^{s+1} \sum_{1<i_1<\cdots<i_s} \frac{e(p_{i_1})\cdots e(p_{i_s})}{p_{i_1}\cdots p_{i_s}} + \sum_{s=t+1}^{\pi(z)} (-1)^{s+1} \sum_{1<i_1<\cdots<i_s} \frac{e(p_{i_1})\cdots e(p_{i_s})}{p_{i_1}\cdots p_{i_s}} \tag{4.5}$$

$$\Updownarrow$$

$$S(A,P,z) \leq N\prod_{p<z}\left(1 - \frac{e(p)}{p}\right) + \sum_{s=0}\binom{t}{s}2^s + N\sum_{s=t+1}^{\pi(z)} (-1)^{s+1} \sum_{1<i_1<\cdots<i_s} \frac{e(p_{i_1})\cdots e(p_{i_s})}{p_{i_1}\cdots p_{i_s}} \tag{4.6}$$

A closer analysis shows that $N\prod_{p<z}\left(1-\frac{e(p)}{p}\right)$ dominates the upper bound in (4.6), so we can simply estimate $N\prod_{p<z}\left(1-\frac{e(p)}{p}\right)$. We claim

$$\prod_{p<z}\left(1-\frac{e(p)}{p}\right)\ll\frac{1}{\log z^2}. \tag{4.7}$$

To see this, note that

$$\prod_{p<z}\left(1-\frac{e(p)}{p}\right)\leq e^{\left(-\sum_{p<z}\frac{e(p)}{p}\right)}\ll\frac{1}{(\log z)^2} \tag{4.8}$$

where the first equation holds because $1-x\leq e^{-x}$, and the second is left to the reader. Finally $\pi_2(x)<S(A,P,z)\leq\frac{N}{(\log z)^2}$, and letting $z=\frac{\log N}{C\log\log N}$ establishes the theorem. □

4.4 Modern Sieves

4.4.1 The Selberg Sieve

As with the Legendre and Brun sieves, many other sieves revolve around carefully choosing which terms to include, and using them to find bounds. These sieves are known as *combinatorial sieves*. Because of its large error terms, the Legendre Sieve is not very useful in practice. Modern sieves include Brun's Sieve, Gallagher's Sieve, the Turan Sieve, and the Selberg Sieve, which resolve the problems of the Legendre Sieve by seeking to bound $S(A,B,z)$, instead of computing it exactly. We now describe the Selberg Sieve, and its most important application. This requires a bit more background on the Mobius function.

Definition 5. An arithmetic function f is a complex valued function defined on the positive integers. A multiplicative function is an arithmetic function such that $\forall a,b$ where $(a,b)=1, f(ab)=f(a)f(b)$. The summation function F of f is an arithmetic function defined by

$$F(n)=\sum_{d|n}f(d).$$

Theorem 6 (Mobius inversion formula). *If f is an arithmetic function and F is the summation function of f then*

$$f(n)=\sum_{d|n}\mu(d)F\left(\frac{n}{d}\right).$$

Proof. Following [4], we have:

$$\sum_{d|n}\mu(d)F\left(\frac{n}{d}\right)=\sum_{d|n}\sum_{c|\frac{n}{d}}f(c)=\sum_{c|n}\sum_{d|\frac{n}{c}}\mu(d)f(c)=\sum_{c|n}f(c)\sum_{d|\frac{n}{c}}\mu(d)=f(n)$$

where the last equality follows from the fact that for $\frac{n}{c}>1, \sum_{d|\frac{n}{c}}\mu(d)=0$. □

Now that we've familiarized ourselves with Mobius inversion, we can state the Selberg Sieve. Let A be a set of positive integers $\leq t$, and P be a set of primes. Note that if d is a product of distinct primes, then A_d is the intersection of the A_p for all $p|d$. Let $P(k)$ denote the product of the primes less than the real number k. As usual $S(A,P,k)$ will denote the set of elements in A that aren't divisible by any prime $p<k$. It turns out that $|A_d|$ can be estimated by

$$|A_d|=\frac{1}{f(d)}|A|+R_d$$

where f is a multiplicative function and R_d is the error term. Let g be the Mobius inversion of f, i.e. $g(n) = \sum_{d|n} \mu(d) f\left(\frac{n}{d}\right)$, and $f(n) = \sum_{d|n} g(d)$. Let $V(k) = \sum_{d<k,\, d|P(k)} \frac{\mu^2(d)}{g(d)}$. Using this estimation and an inclusion-exclusion argument similar to the one we gave for Brun's Sieve we have an upper bound due to Selberg:

Theorem 7 (Selberg Sieve).

$$S(A,P,k) \leq \frac{|A|}{V(k)} + O\left(\sum_{d_1,d_2<z,\, d_1,d_2|P(z)} |R_{[d,d_2]}|\right)$$

Among its many other uses the Selberg Sieve can be used to establish Brun's Theorem, and most notably:

Theorem 8 (Brun-Titchmarsh Theorem). *If $S(x,a,q)$ denotes the number of primes $p \leq x$ congruent to a modulo q then*

$$S(x,a,q) \leq \frac{2x}{\phi(q)\log(\frac{x}{q})}$$

4.4.2 The Future of Sieve Theory

Like many branches of mathematics, sieve theory evolved in pursuit of a few legendary problems, namely the Goldbach Conjecture and the Twin Prime Conjecture. The Goldbach conjecture states that every integer greater than 2 can be written as the sum of two primes. The Twin Prime Conjecture is that there are infinitely many twin primes. Sieve Theory has managed to approximate these problems in incredible ways; in 1966 Chinese mathematician Chen Jingrun proved that every sufficiently large even number can either be written as the sum of two primes or a prime and a number that is the product of two primes, and as we've seen Brun showed the sum of the reciprocals of the twin primes converge. Another strong Sieve Theory result is the Friedlander-Iwaniec Theorem which states that there are infinitely many primes of the form $a^2 + b^4$. While Sieve Theory techniques do seem powerful, they are limited by what is known as the "parity problem": sieve methods can't distinguish between numbers with an even number of prime factors, and numbers with an odd number of prime factors. Even the formidable Terence Tao who used sieve methods to prove his Tao-Green theorem on primes in arithmetic progression despairs on his blog that "it is probably premature with our current understanding to try to find a systematic way to get around the parity problem in general but it seems likely that we can get around it in some cases."

References

[1] A.C. Cojocaru and M.R. Murty, *An Introduction to Sieve Methods and their Applications*, London Mathematical Society Student Texts, **66**, Cambridge University Press 2005.

[2] H. Halberstam and H.E. Richert, *Sieve Methods*, Academic Press, 1974.

[3] M.S.H. Faester, *Brun's Theorem and the Sieve of Eratosthenes*, available at `http://www.cs.au.dk/~mshf`.

[4] T. Andreescu, D. Andrica, and Z. Feng, *104 Number Theory Problems: From the Training of the USA IMO Team*, Birkhäuser, 2007.

[5] T. Tao, *Open Question: The parity problem in sieve theory*, available at `http://terrytao.wordpress.com`, 2007.

FACULTY FEATURE ARTICLE

5

A Direct Geometric Proof of Gregory's series for $\frac{\pi}{4}$

Paul G. Bamberg[†]
Harvard University
Cambridge, MA 02138
bamberg@math.harvard.edu

Abstract

In the spirit of a proof that was done in 15th century India long before the birth of James Gregory, and without any explicit mention of the arc tangent function, Gregory's series is shown by a geometric argument to sum to one-fourth the area of the unit disc.

5.1 Introduction

One of the most famous formulas involving π is the infinite series known as "Gregory's series:"

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots .$$

The usual derivation of this series relies on the differentiation rule for the arc tangent function:

$$\frac{d}{dx} \arctan x = \frac{1}{1+x^2},$$

from which it follows that

$$\frac{\pi}{4} = \arctan 1 = \int_0^1 \frac{1}{1+x^2}\, dx.$$

According to the binomial expansion,

$$(1+x^2)^{-1} = 1 - x^2 + x^4 - x^6 + \cdots ,$$

so

$$\frac{\pi}{4} = \int_0^1 (1 - x^2 + x^4 - x^6 + \cdots)\, dx = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots .$$

This formula was known to the great 15th century mathematicians of Kerala, India, notably Nilakantha and Jyestadeva, who wrote about it in Sanskrit verse [1], but they could not have known it as Gregory's series, for the simple reason that James Gregory (1638–1675) had not been born yet. Neither, for that matter, had Newton, inventor of differential calculus and of the binomial theorem for arbitrary exponent, both needed for the standard derivation. Furthermore, the standard derivation makes no contact with a geometrical definition of π either in terms of the area or circumference of a circle.

[†]Paul Bamberg studied physics at Harvard and Oxford, taught physics at Harvard from 1967 to 1995, and returned to Harvard in 2000 as a member of the mathematics department. Between 1980 and 2000, as a founder of Dragon Systems, he invented and implemented algorithms for speech recognition on personal computers, but his first love has always been teaching.

My purpose is to present a simple derivation of Gregory's series that defines π as the area of the unit circle, that uses no differentiation formulas except for the definition of the derivative, and that never mentions infinite series. It was inspired by a proof first done by Jyestadeva [1], but every detail has been changed. In particular, Jyestadeva set out to calculate the length of a circular arc, while I have used an "area squeeze" argument.

5.2 A Geometric Proof

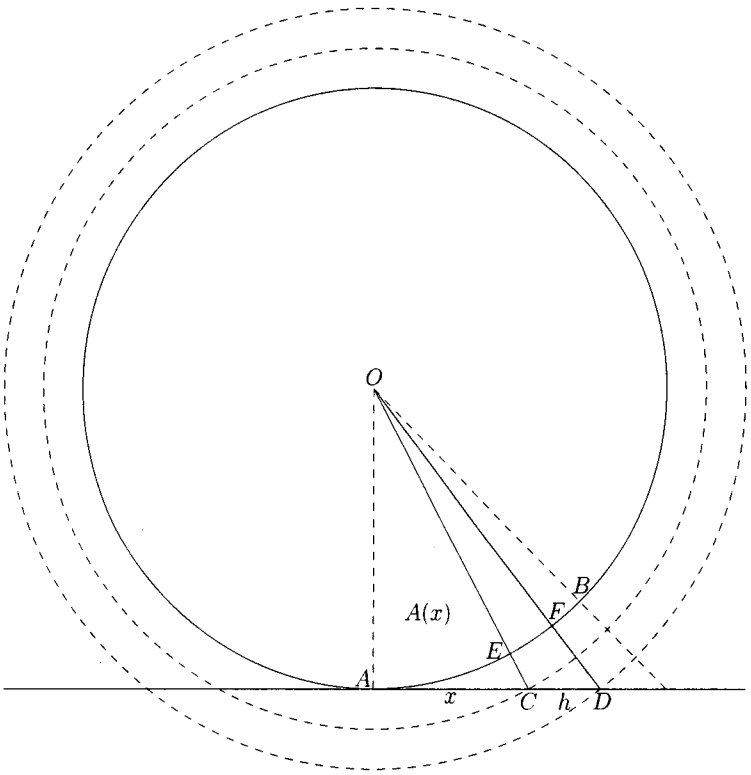


Figure 5.1: Area Squeeze Diagram

Everything is based on Figure 5.1. The circle through A and B has a radius of 1 and an area of π. The radius OB intersects the horizontal axis at $x = 1$, and so the wedge AOB has area $\frac{\pi}{8}$.

Regard the area of a circular wedge as a function of x-coordinate of the point where the radius intersects the x-axis. Thus, in the diagram, $A(x)$ is the area of wedge AOE, while $A(x+h)$ is the area of wedge AOF.

The area of a wedge scales as the square of the radius, so the wedge EOF, when scaled up so that it is bounded by a circular arc through C, has area $(A(x+h) - A(x))(1+x^2)$. The same wedge, when scaled up so that it is bounded by a circular arc through D, has area $(A(x+h) - A(x))(1+(x+h)^2)$.

Look at the triangle COD, whose area is $\frac{h}{2}$. This triangle contains the smaller wedge (through C) and is contained in the larger wedge (through D). Since the area of a wedge is a continuous

function of its radius, it follows that there is a value H, satisfying $0 < H < h$, such that

$$\frac{h}{2} = \big(A(x+h) - A(x)\big) \cdot \big(1 + (x+H)^2\big).$$

Now we can calculate the derivative $A'(x)$ directly from its definition:

$$A'(x) = \lim_{h \to 0} \frac{A(x+h) - A(x)}{h} = \lim_{h \to 0} \frac{1}{2(1 + (x+H)^2)}.$$

As $h \to 0$, so does H. We conclude that

$$A'(x) = \frac{1}{2(1+x^2)}.$$

It follows that

$$\frac{\pi}{4} = 2A(1) = \int_0^1 \frac{dx}{1+x^2}.$$

Once you understand the result, it makes intuitive sense. If you evaluate the integral $\int_0^1 dx$, you get the area of a triangle with base 1 and height 1. If, instead, you want the area of the circular wedge, you merely scale down the integrand by a factor of $\frac{1}{1+x^2}$.

How, now, do we proceed without the binomial expansion? First observe that

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

Iterating n times, we get the exact result

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \cdots \pm x^n \frac{1}{1+x^2}.$$

Integrating term by from 0 to 1 (Jyestadeva found this hard to do), we obtain

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots \pm C \cdot \frac{1}{2n+1},$$

where, since $0 < \frac{1}{1+x^2} < 1$ over the interval of integration, C must lie between 0 and 1.

References

[1] R. Roy. *The discovery of the series formula for π by Leibniz, Gregory and Nilakantha.* Math. Mag., (1990) 63(5), 291–306.

FEATURE

6

MATHEMATICAL MINUTIAE

$1 + 1 = 1$?

Katrina Evtimova[†]
Harvard University '13
Cambridge, MA 02138
kevtimova@college.harvard.edu

Can the "identity" $1 = 1 + 1$ make any mathematical sense? While most people would instantly say, "No way!," it turns out that under certain conditions, the answer to this question is positive. Indeed, although this strange "identity" may seem quite contradictory at first, there are rigorously proved results in mathematics which favor it. For example, such a mathematical paradox was proved by the Polish mathematicians Stefan Banach and Alfred Tarski in the 1920s and can be stated as follows:

Theorem 1 (Banach-Tarski Paradox). *A solid sphere in $\mathbb{R}^3$ can be divided into finitely many pieces which, using only rigid motions, can be reassembled to form two distinct spheres of the same volume as that of the original one.*

In this article, we are going to outline a rough idea of why this theorem holds, following the exposition in [3].

But let us first consider an example that will make us prone to believe that $1 = 1 + 1$ may actually hold. Let P be the set of all polynomials in $e^{i\varphi}$ with nonnegative integer coefficients. We partition P into two non-intersecting subsets, A and B, where A is the set of all polynomials in P with a zero constant term and B is the set of all polynomials in P with a nonzero constant term. We have that P is the disjoint union of A and B. We will show that using only rigid motions, i.e., those motions of the plane that preserve relative distances between points, we can get P from A and from B separately.

Indeed, if we multiply all elements in A by $e^{-i\varphi}$, i.e., rotate A by φ radians clockwise, we obtain the whole set P. Also, if we add the constant 1 to the elements of B, i.e., translate B by 1, we again obtain the whole set P. Thus we showed that P is decomposable into two copies of itself. Quite amazing! This helpful example is an illustration of the Sierpinski-Mazurkiewicz Paradox. The reader is encouraged to learn more about this mathematical phenomenon in [2] and [3].

Now, let's turn to the proof of the Banach-Tarski Paradox. How is it possible to decompose a solid sphere into finitely many pieces that, when reassembled, give two distinct spheres of the same volume as the original one? The key points are the properties of the irrational numbers and the Axiom of Choice.

For simplicity, imagine that the Earth is an ideal solid sphere. Consider two axes of rotation for the Earth — the first one (l_1) with endpoints being the north and south poles and the second one (l_2) with endpoints being the two intersections of the Greenwich meridian with the equator.

Let θ be an irrational number. We denote by r_1 the clockwise rotation of θ degrees around l_1 and by r_2 the clockwise rotation of θ degrees around l_2. Let r_1^{-1} and r_2^{-1} be the inverse rotations of r_1 and r_2, respectively. We define S to be the free group generated by $r_1, r_2, r_1^{-1}, r_2^{-1}$ where the rotations in the pairs (r_1, r_1^{-1}) and (r_2, r_2^{-1}) cancel each other. In other words, S is the set of all finite strings of rotations r_1, r_2, r_1^{-1}, and r_2^{-1} such that no string has the rotations r_1 and r_1^{-1} or r_2 and r_2^{-1} adjacent to each other in its expression. Each element of S can be considered as a

[†]Katrina is a sophomore in Mather house concentrating in Math. She is an international student from Bulgaria who is currently most interested in abstract algebra. In the past, Katrina has conducted research in representation theory.

sequence of rotations which are applied from the left to the right. Note that by the irrationality of θ, the orders of elements in S are infinite, since there is no positive integer n such that after applying either of the rotations r_1 or r_2 exactly n times, we end up at the point that we started with. Namely, for a point P, we have that $r_i^n(P) \neq P$ since otherwise, it would follow that $n\theta$ is a multiple of 360°, i.e., that θ is rational. Therefore, there are infinitely many elements in S.

In order to keep things simple, we will consider only the points on the surface of the Earth. Let us denote them by the set E. The following is a sketch of an algorithm for partitioning E into smaller disjoint subsets. Note that the effectiveness of this algorithm is not rigorously proved, as our aim is just to get a sense of the general idea of the proof. Let A_1 be a point in E. Let E_1 be the subset of E obtained by applying all elements in S to A_1. We choose a point A_2 in E such that A_2 is not in E_1. Such a point exists since the number of points in S is countably infinite, whereas the number of points in E is uncountably infinite. Again, we apply all elements of S to it and continue in the same manner. Thus, we obtain the disjoint sets $E_1, E_2, \ldots \subset E$. We need uncountably many points A_i in order to cover all points in E by the union of all E_i since a countable union of countable sets is still countable but E is uncountable.

Now, by the Axiom of Choice, we can choose a unique point from each of the subsets E_i and thus construct a new subset D of E. By the way D is constructed, we conclude that every point in E can be covered by applying a unique element of S to a unique point in D. In other words, each point in E has a unique representation as an element of S applied to a point in D.

Let E_{r_1} be the subset of E of points obtained by applying an element of S, starting with the rotation r_1, to some point in D. We define E_{r_2}, $E_{(r_1)^{-1}}$, and $E_{(r_2)^{-1}}$ in a similar way. The last major point in our outline is to note that by applying the rotation r_1^{-1} to the points in E_{r_1}, we obtain all the points in $E_{r_1} \cup E_{r_2} \cup E_{(r_2)^{-1}}$. Similarly, applying r_2^{-1} to E_{r_2} gives all the points in $E_{r_1} \cup E_{r_2} \cup E_{(r_1)^{-1}}$. Therefore, the set E can be partitioned in the following two ways:

$$E = r_1^{-1}(E_{r_1}) \cup E_{r_1^{-1}},$$
$$E = r_2^{-1}(E_{r_2}) \cup E_{r_2^{-1}}.$$

These two decompositions imply that starting with the set E, we can partition it into 4 subsets from which, using only rigid motions, we can construct two sets E. This result is very similar to the Banach-Tarski paradox, but we are still missing a few important points which we are just going to state. The first one is that we restricted ourselves to the surface of the Earth. We need to extend the outlined algorithm to the whole volume of the Earth. The second one is that when we extend, it there are some points that are left out by our algorithm, namely the points on the two axes of rotation l_1 and l_2 and the points in D. In the end, it turns out that the Earth can be divided into exactly 5 pieces which, when reassembled, give two spheres of the same volume as that of the Earth. This is exactly what Banach and Tarski proved. As mentioned above, what we outlined so far is just the idea of why 1 holds. If you are interested in a more rigorous proof of this fact, you are encouraged to consult [3].

But you should not think that you can actually perform the decomposition just sketched. It is not physically possible since matter is not infinitely divisible. In spite of that, this counterintuitive result is a brilliant example of the beauty of mathematics and its power to challenge our imagination.

References

[1] F. Su. *Banach-Tarski Paradox.* Math Fun Facts.
http://www.math.hmc.edu/funfacts/ffiles/30001.1-3-8.shtml

[2] F. Su. *Sierpinski-Mazurkiewicz Paradox.* Math Fun Facts.
http://www.math.hmc.edu/funfacts/ffiles/30001.1-2-8.shtml

[3] F. Su. *The Banach-Tarski Paradox.*
http://www.math.hmc.edu/~su/papers.dir/banachtarski.pdf

FEATURE

7

STATISTICS CORNER

Random Walk Model For Dating

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

With the food on my tray, I started the mundane process of finding a dinner buddy again. But what caught my eyes instead was a beautiful girl with big eyes and sweet lips sitting alone. Taking down a shot of confidence, I put my tray down right across from her and introduced myself...

What are the chances for us to be together?

For the sake of simplicity, let's assume the following:

- We meet everyday and hang out.
- Each day I have a probability p of making her happier and q of making her less happy, where $p+q=1$. Similarly she has a probability r of making me happier and s of making me less happy, where $r+s=1$.
- We can measure the happiness of each person with the other by an integer. On day 0, each person has happiness 0. Every day after that, the happiness of each person either decreases or increases by 1, with the probability described above.
- Suppose when I have happiness C (a positive integer), then I'm ready for a relationship, and correspondingly define D for her. Thus, only when I have more happiness than C and when she has more happiness than D will we date.

Then we can map our "Happiness State" with an ordered pair $(P,K)_i$ at day i on a 2-dimensional plane. Let $U[(P,K)_i]$ be the probability that this happiness state is (P,K) on day i. We have the following equations, given the above assumptions:

$$\begin{aligned} U[(0,0)_0] &= 1, \\ U[(x,y)_{i+1}] &= prU[((x-1),(y-1))_i] + qrU[((x+1),(y-1))_i] \\ &\quad + psU[((x-1),(y+1))_i] + qsU[((x+1),(y+1))_i]. \end{aligned}$$

Thus now we ask (slightly qualifying the initial question): What's the chance of us dating at time t? In math terms, what's the probability of (P,K) with $P \geq C$ and $K \geq D$ at t? In other words, we seek to find

$$\sum_{P \geq C, K \geq D} U[(P,K)_t].$$

Theorem 1. *Given that a particle starts at 0 on the number line at time 0, and that at each second it can either move left or right by 1, the number of ways to arrive at position k at time t is $\binom{t}{(t-k)/2}$.*

†Greg Yang, Harvard '14, plans to concentrate in mathematics. He has much experience from High School Math Olympiads in solving difficult problems. Since coming to Harvard, he has begun to explore interesting applications of math to social sciences.

Proof: At time t, the particle has taken t steps, each either left or right (denoted L or R respectively). We may represent the sequence of steps with a sequence of letters L and R. If $k > 0$, then $(t - |k|)/2$ of them are L; if $k \leq 0$, then $(t - |k|)/2$ of them are R. Then we just need to choose where those $(t - |k|)/2$ letters appear from the total t letters, which is $\binom{t}{(t-|k|)/2}$. This is also $\binom{t}{(t+|k|)/2}$, so one of them must be $\binom{t}{(t-k)/2}$. □

Corollary 2. *The polynomial $(x+y)^t$ generates the above sequence, with the coefficient of $x^a y^{t-a}$ being the number of ways to arrive at position $(2a - t)$.*

If we extend the result of this corollary to our 2 dimensional grid, and observe that vertical movement happens simultaneously but independently from horizontal movement, we get the following:

Corollary 3. *In $(x + y)^t(n + m)^t$, the coefficient of $x^a y^{t-a} n^b m^{t-b}$ corresponds to the number of ways to get to $(2a - t, 2b - t)$.*

Corollary 4. *Define the polynomial*

$$E_t(x, y, n, m) := \sum_{2a-t \geq C, 2b-t \geq D} x^a y^{t-a} n^b m^{t-b} ([x^a y^{t-a} n^b m^{t-b}](x + y)^t (n + m)^t),$$

where $[x^k]$ represents the coefficient extractor of the polynomial succeeding it. Then $E_t(p, q, r, s)$ is the probability that we are dating at time t.

Notice that

$$E_t(x, y, n, m) = \left(\sum_{2a-t \geq C} x^a y^{t-a} [x^a y^{t-a}](x + y)^t \right) \cdot \left(\sum_{2b-t \geq D} n^b m^{t-b} [n^b m^{t-b}](n + m)^t \right).$$

Let $F_t(x, y) = \sum_{2a-t \geq C} x^a y^{t-a} [x^a y^{t-a}](x + y)^t$. If we use F_t to represent $F_t(p, q)$ for the sake of simplicity, then with a little algebra we can show that

$$F_{C+2k+2} = \sum_{h=0}^{k} \binom{C + 2h + 1}{C + h} p^{C+h+1} q^{h+1} - \sum_{h=0}^{k} \binom{C + 2h}{C + h} p^{C+h} q^{h+1} + F_C$$

where $F_C = p^C$.

This series is easily calculable with Mathematica for finite k. But what happens as k goes to infinity? In other words, what's the chance that she is willing to date me forever? (I hope it's 1.)

Let $S_C = \sum_{h=0}^{\infty} \binom{C+2h+1}{C+h} p^{C+h+1} q^{h+1} - \sum_{h=0}^{\infty} \binom{C+2h}{C+h} p^{C+h} q^{h+1} + p^C$. Let's first investigate the case when $C = 1$. Then

$$S_1 = \sum_{h=0}^{\infty} \binom{2h + 2}{1 + h} p^{h+2} q^{h+1} - \sum_{h=0}^{\infty} \binom{1 + 2h}{1 + h} p^{1+h} q^{h+1} + p.$$

Notice that $\sum_{h=0}^{\infty} \binom{1+2h}{1+h} p^{1+h} q^{h+1} = \frac{1}{2} \sum_{h=0}^{\infty} \binom{2+2h}{1+h} p^{1+h} q^{h+1}$. But $\mathcal{C}(x) = \sum_{i \geq 0} \binom{2i}{i} x^i/(i+1)$ generates the Catalan numbers $C_{n+1} = \sum_{k=0}^{n} C_k C_{n-k}$. This recursive relation gives $\mathcal{C} = 1 + x\mathcal{C}^2$, which by the quadratic formula implies $\mathcal{C}(x) = \frac{1-\sqrt{1-4x}}{2x}$. When $\mathcal{C}$ is treated as a Taylor expansion around 0, it has convergence radius $1/4$. (This is a brief treatment of the Catalan numbers and their generating function. For more information, consult combinatorics textbooks.)

Note that $\frac{d}{dx}(x\mathcal{C}) = \frac{d}{dx}(\frac{1-\sqrt{1-4x}}{2}) = (1 - 4x)^{-1/2}$. On the other hand,

$$\frac{d}{dx}(x\mathcal{C}) = \frac{d}{dx} \sum_{i \geq 0} \binom{2i}{i} x^{i+1}/(i + 1) = \sum_{i \geq 0} \binom{2i}{i} x^i.$$

This sum is convergent for $x \in (-1/4, 1/4)$. If $p \neq 1/2 \Rightarrow pq = p(1-p) < 1/4$, we find

$$\begin{aligned} S_1 &= p\left(\frac{d}{dx}(xC)(pq) - 1\right) - \frac{1}{2}\left(\frac{d}{dx}(xC)(pq) - 1\right) + p \\ &= \left(p - \frac{1}{2}\right)\left((1-4pq)^{-1/2} - 1\right) + p \\ &= \left(p - \frac{1}{2}\right)(1 - 4p(1-p))^{-1/2} + \frac{1}{2} \\ &= \left(p - \frac{1}{2}\right)/|1-2p| + \frac{1}{2} \\ &= \frac{1}{2}\left(\operatorname{sgn}\left(p - \frac{1}{2}\right) + 1\right) \\ &= \begin{cases} 1 & \text{if } p > 1/2, \\ 0 & \text{if } p < 1/2. \end{cases} \end{aligned}$$

In the case of $p = q = 1/2$, we may just use the definition of F_t:

$$\begin{aligned} \lim_{t\to\infty} F_t &= \lim_{t\to\infty} \sum_{2a-t\geq C} \binom{t}{a} p^a q^{t-a} \\ &= \lim_{t\to\infty} \sum_{2a-t\geq C} \binom{t}{a} (1/4)^t \\ &= \lim_{t\to\infty} \frac{1}{2}\left(1 - \sum_{t-C<2a<C+t} \binom{t}{a}(1/4)^t\right). \end{aligned}$$

However, by Stirling's approximation,

$$\begin{aligned} \lim_{n\to\infty} \binom{2n}{n}(1/4)^{2n} &= \lim_{n\to\infty} \frac{\sqrt{4\pi n}(2n/e)^{2n}}{(\sqrt{2\pi n}(n/e)^n)^2}(1/4)^{2n} \\ &= \lim_{n\to\infty} \frac{4^n}{\sqrt{\pi n}}(1/4)^{2n} \\ &= 0. \end{aligned}$$

And because $\binom{2n}{n} \geq \binom{2n}{k}$ for all k, we have $0 = \lim_{n\to\infty}\binom{2n}{n}(1/4)^{2n} \geq \lim_{n\to\infty}\binom{2n}{k}(1/4)^{2n} \geq 0$, so we actually have equalities here. Similarly,

$$\lim_{n\to\infty} \binom{2n-1}{n}(1/4)^{2n-1} = \lim_{n\to\infty} \frac{1}{2}\binom{2n}{n}(1/4)^{2n}(4) = 0.$$

And $0 = \lim_{n\to\infty}\binom{2n-1}{n}(1/4)^{2n-1} \geq \lim_{n\to\infty}\binom{2n-1}{k}(1/4)^{2n-1} \geq 0$ for all k, so we actually have equalities here as well.

Then $\lim_{t\to\infty}\sum_{t-C<2a<C+t}\binom{t}{a}(1/4)^t = 0$, as the sum has a finite number (at most C) of terms, and each of these terms goes to 0. Hence $\lim_{t\to\infty}(1-\sum_{t-C<2a<C+t}\binom{t}{a}(1/4)^t)/2 = 1/2$ is the probability that she is willing to date me forever, if $p = q = 1/2$. (Well, at least it's pretty fair.)

We can use the same strategy above to prove that S_C is constant regardless of C. In other words, $\lim_{t\to\infty} F_t$ is the same for any value of C.

This is a surprising result. It means that if the girl makes me happier in more than half of the days, then I am willing to date her forever; if she makes me happier in less than half of the days, then I would not want to date her in the long run. Applying the same logic on her, it follows that we will date (and probably be married) eventually if and only if each of us makes the other happier in more than half of the time.

We can generalize this model by changing the time interval that corresponds to t (for example, change "day" to "hour"). This model may even apply to other social situations that involve quasi-random interactions such as friendships.

Now we must consider the flaws of this random walk model used for a dating scenario.

1. Even though one cannot predict the effects of one's actions on another, he or she still has a choice of whether to show signs of affection or annoyance. This model completely ignores choice.

2. The behaviors of the two are different in different stages of their relationship. So perhaps the p and r should vary depending on the happiness state.

3. There are many other social factors influencing the two. By only considering the interactions between them and not between them and everyone else, the model is ignoring effects of things like peer pressure and multiple suitors on the happiness state.

4. The practicality of the model still rests on the knowledge of one's p value, which is not easily extracted.

5. We should consider different degrees of happiness. So perhaps we should use a probability distribution more complex than the dichotomy presented above.

6. And more.

But if I neglect these flaws for a second, it seems that we are either destined to be together or...

"May I sit here?"
"Oh sorry, my boyfriend is sitting there."

...destined to be apart.

FEATURE

8

APPLIED MATHEMATICS CORNER

A Novel Dual-Layered Approach to Geographic Profiling in Serial Crimes

Zhao Chen[†]
Harvard University '09
Cambridge, MA 02138
zhaochen@fas.harvard.edu

Kevin Donoghue[‡]
Harvard University '09
Cambridge, MA 02138
kdonogh@fas.harvard.edu

Alexander Isakov[*] (corresponding author)
Harvard University '09
Cambridge, MA 02138
isakov@fas.harvard.edu

Abstract

The formalized theory of geographic profiling has advanced substantially, both theoretically and in terms of wide-spread practical use, since its beginnings in 1989. In this paper, we propose a novel model for geographic profiling, using a dual-layered approach that incorporates more than simply spatial considerations to predict future criminal behavior. We implement our model by generating a three-dimensional plot laid upon a city grid that plots "danger zones" where a serial criminal might next strike. We test our model against intuitively obvious cases, and follow up by applying the model to the case of the arsonist Mr. Thomas Sweatt. Finally, we suggest some natural extensions and variants for further research and implementation.

8.1 Introduction

This paper on geographic profiling is a modified version of our submission to the Mathematical Contest in Modeling, a four-day international competition where teams of up to three people propose an original model to a given real-world problem. Our team was ranked Finalist for a discrete model that can be used for catching serial criminals.

Given a serial criminal, the place and time of his or her crimes can be used to generate a prediction of either where the criminal resides or where the next crime will be committed. *Such*

[†]Zhao Chen is a senior in Pforzheimer House at Harvard College concentrating in Physics and Mathematics. His primary experimental interests lie in atomic/molecular/optical physics and condensed matter, but as a mathematical hobbyist he also enjoys graph theory and algebra. His other interests include food and East Asian history/literature.

[‡]Kevin Donoghue is a senior at Harvard concentrating in Mathematics. Currently, he is mostly concerned with geometry and analysis, although he enjoys all kinds of pure math. He is also interested in classical music.

[*]Alexander Isakov is a senior at Harvard concentrating in Physics and Mathematics. His primary interests are physics generally, and nonlinear dynamics in particular, but as a hobby he enjoys programming and mathematical modeling with friends. He is also interested in the classics.

predictions are known as geographic profiles, and they have been used in a variety of cases, such as the one of the serial arsonist Thomas Sweatt [7], whose crime map [18] is shown in Fig. 1.

Indeed, as geographic profiling becomes more prevalent in investigative strategies, the need for more accurate methods of modeling criminal activity has grown. Preexisting models rely heavily on geographic data [10], and as such do not take fully into account the crime information available. We have found that it is possible to prevent such waste of information, and create a more sophisticated profiling system which incorporates more of the available data.

8.1.1 Problem Background

Serial criminals exhibit certain pathologies that allow predictive measures to be taken against them. Most importantly, in contrast to one-time criminals, serial criminals tend to target strangers, which makes their actions more prone to patterns [11]. Beginning with the seminal work of Brantingham and Brantingham in 1981 [2], criminologists began to turn to geographic profiling as a useful complement to crime investigations [11]. According to Brantingham and Brantingham, serial criminals tended to operate within a relatively limited "activity space." Fueled by these ideas, investigators began to use the geographic and temporal information about committed crimes to generate a probabilistic map that helped them predict where the criminal resided. The majority of models used various spatial metrics to define a "geographic center" of the crimes committed, and predicted that the criminal resided in close proximity of this center. Such methods were often very simple, using basic mathematical operations to determine this central point [12].

Kim Rossmo, a student of the Brantinghams, attempted to modify this model in 1987 with a "buffer zone," reasoning that the criminal would not be too comfortable committing crimes in a certain radius around his or her home [13]. In later work, Rossmo's model was further refined to include multiple "anchor points" (as opposed to a single point) where the suspect was predicted to live. In other work, location of body dump sites were taken into consideration by looking at how far a suspect would be able to take the body [12]. Despite such refinements, however, Rossmo has admitted that these methods are far from perfect and must always be used in tandem with traditional tried-and-true methods [13]. Many studies have shown educated individuals consistently coming up with just as accurate geographical predictions as the most sophisticated modeling software. [5, 14]

Moreover, despite the effort put into geographic profiling as a tool to locate criminals, relatively little attention has been given to predicting an uncaught serial criminal's next target. Given the wealth of information on how serial criminals think and plan out their crimes, this type of crime prediction holds much promise.

We propose a new model to be used for geographic profiling. After testing our model against a series of limiting cases, we use it to determine the correct place of residence and a very accurate danger zone in the case of the Washington D.C. arsonist Thomas Sweatt. We show that from the data of the first ten Sweatt arson locations, our model accurately predicts the neighborhoods that are affected by later crimes as well as the residence of the arsonist. Indeed, one of the neighborhoods that the model designates as most likely to contain the criminal actually does contain Mr. Sweatt's residence.

The paper is organized as follows. First, we detail the model. Then, we describe how our model performs with a fictitious criminal who exhibit distinct behavioral patterns. We also provide a thorough treatment of the real-life Thomas Sweatt case. Finally, we conclude and suggest possible opportunities for further research.

8.2 The Model

8.2.1 Definitions and Terms

Our model will be based on an $N \times N$ grid called a city. Every point on the grid is identified by its unique Cartesian coordinate (x, y), and is known as a neighborhood. Each neighborhood is assigned parameters such as densities of different races, average wealth, etc., as shown in Table 1.

Throughout this paper, we deal with serial criminals and glean information from the crimes they commit. A crime refers to a certain cartesian coordinate (x, y) and an index i that orders the crimes *by time, as well* as information about the neighborhood and the particular victim (race, gender, age,

average wealth of neighborhood, average safety of neighborhood, etc.). Such information together is called the crime vector. Every crime that the serial criminal commits has its own associated crime vector. The crime vector of the ith crime is denoted $\mathfrak{C}_i$, and its different components are denoted $\mathfrak{C}_i$.component. For example, if the third crime were committed against race R_A, then $\mathfrak{C}_3.R = R_A$. If this crime were committed at the neighborhood $(3,3)$, then $\mathfrak{C}_i.x = \mathfrak{C}_i.y = 3$. For ease of notation, we denote the location of the ith crime as the point $C_i = (\mathfrak{C}_i.x, \mathfrak{C}_i.y)$.

The metric we use for distance in our model is a modified version of the Manhattan Distance, which was used in Rossmo's original model [12]. The Manhattan Distance between (x, y) and (x', y') is $|x - x'| + |y - y'|$. However, we also account for more efficient transportation available (highways, subways, bus routes, etc.) by saying that the distance between two points $p_1 = (x, y)$ and $p_2 = (x', y')$ connected by fast transportation is $\alpha(|x - x'| + |y - y'|)$, where $0 < \alpha < 1$. We use $\alpha = \frac{1}{3}$ and all points with coordinate $x = 3$ are connected by public transportation. Hence, we define the distance between two points $D(p_1, p_2)$ as the minimum distance of any path on our grid leading from p_1 to p_2. In our test city, where fast transportation lies along the line $x = 3$, this is precisely:

$$D(p_1, p_2) = \min\{|x - x'| + |y - y'|, |x - 3| + |x' - 3| + \alpha|y - y'|\}. \tag{8.1}$$

The goal is to generate a danger map which is our generated grid along with a value (by our algorithm, this value will be between 0 and 2) associated with each cell that is correlated with the probability of the serial criminal's next crime being committed there. We do this by generating a psychological danger mapping and a spatial danger mapping, and superimposing these two maps with appropriate weighting. More precisely, for a grid of size $k \times \ell$, we define two mappings

$$\delta_p : \{0, \ldots, k-1\} \times \{0, \ldots, \ell - 1\} \to \mathbb{R} \tag{8.2}$$

$$\delta_s : \{0, \ldots, k-1\} \times \{0, \ldots, \ell - 1\} \to \mathbb{R} \tag{8.3}$$

and call $\delta_p(a, b)$ the psychological danger at the neighborhood (a, b), and $\delta_s(a, b)$ the spatial danger at point (a, b). We also look for the psychological weight ϕ_p and the spatial weight ϕ_s (both real numbers between 0 and 1 inclusive), and we construct the map

$$\Delta = \psi(\phi_p)\delta_p + \psi(\phi_s)\delta_s \tag{8.4}$$

where

$$\psi(x) = \frac{1 + e^4}{e^4 - 1}\left(\frac{1}{1 + e^{-2(k(x - \frac{1}{2}))}} - \frac{1}{1 + e^4}\right) \tag{8.5}$$

This Δ is the total danger mapping, and we call $\Delta(a, b)$ the total danger at the neighborhood (a, b)[1]. This total danger will give us a quantitative measure of how likely a serial criminal will next commit a crime.

Table 1 shows the relevant parameters that will be assigned to each neighborhood in our model.

[1]The function ψ in this expression is there to reduce noise. k in ψ is an empirically determined constant.

List of Crime Independent Parameters			
Parameter Name	**Symbol**	**Meaning**	**Comments**
Race R_i Density	$\rho_i^{(R)}$	The fraction of residents belonging to race R_i in a neighborhood.	For our generated city, there are 4 races, R_A, R_B, R_C, and R_D.
Gender G_i Density	$\rho_i^{(G)}$	The fraction of residents belonging to gender G_i in a neighborhood.	For our generated city, there are 2 genders, G_M and G_F.
Age A_i Density	$\rho_i^{(A)}$	The fraction of residents belonging to age group A_i in a neighborhood.	For our generated city, the age distribution is $A_Y = 15\%$, $A_M = 65\%$, $A_O = 20\%$.
Total Population	P	Total population of a neighborhood.	P is correlated with a neighborhood's position in the city.
Average Wealth	ω	The average wealth of a resident of a neighborhood.	w can be measured by land value or average income.
Average Safety	σ	The average safety of a resident of a neighborhood.	σ can be measured by crime rate.

Table 1: List of Crime Independent Parameters.

We now present the set of crime dependent parameters associated with each neighborhood. By central point, we mean a point p_c such that $\langle D(p, C_i) \rangle$ (the average distance between p and the crime scenes) is minimized. Note that such a point need not be unique; in this case, we allow the computer to randomly choose one of the minima as the center point[2].

List of Crime Dependent Parameters for a Neighborhood p			
Parameter Name	**Symbol**	**Meaning**	**Comments**
Distance to ith crime.	$D(p, C_i)$	The distance from the ith crime to the neighborhood.	The metric for distance is as defined above.
Distance from central point.	$D(p, p_c)$	The distance from the central point to the neighborhood.	None.

Table 2: List of Crime Dependent Parameters.

Because every one of these parameters is defined for a neighborhood, they are all functions of (x, y). Hence, we notate the racial density of race R_A at point (a, b), for example, as $\rho_A^{(R)}(a, b)$. The parameter D_i is also dependent on the crime (the value of i in C_i).

8.2.2 The Urban Grid

We consider a double-tiered crime prediction model that maximizes the use of information available to law enforcement authorities. Clearly, information about crimes encompasses not only the basic location/date/time of crime, but also the type of crime and victim characteristics, such as gender. Whereas many current models primarily use the most basic available information about specific crimes, such as place and time, to predict a base of operations [5] and hence "hot zones", it is more useful to intelligently weight all relevant available information in predicting potential targets. After all, neighborhoods clearly have individual flavor, and the salient characteristics of a crime are not immediately obvious, since there is a strong temptation to guess at psychological

[2]In such a case, this random choice will not distort the results to an appreciable degree. In any situation in which there are two or more minima, the distance on our map between these minima will be small.

motivation and use such guesses in predictions [10]. Insofar as we consider serial crimes [1] that neither target specific individuals (as opposed to contract killings or political abductions) nor specific neighborhoods for their peculiar "local flavor", it is best to first strip the city of all except the most relevant factors that can lead to accurate predictions. We generate a city as follows.

1. **Specify a grid size.** We use a 10×10 grid, which allows for a wide mix of characteristics to be dispersed through the city.

2. **Specify P for each neighborhood.** We assume that population is denser around easy-to-access places, so we choose from a uniform distribution based on distance from the line $x = 3$. We have the population for each neighborhood drawn from $\lfloor \text{Unif}(10, 100-10(|x-3|)) \rfloor$, so that each neighborhood has at least 10 people.

3. **Specify w and σ for each neighborhood.** These relative numbers will be useful in gauging a criminal's psychological parameter. We suppose that on average wealthier and safer neighborhoods are farther from the city center. σ, ω are drawn from $\lfloor \text{Unif}(1, 2 + 1/2(|x - 4| + |y - 4|)) \rfloor$, which will avoid unreasonable disparities in wealth and safety.

4. **Input transporation.** Adjacent points connected by fast transportation are counted as only $\frac{1}{3}$ units of distance apart. This is important since serial criminals often take transportation into account when they plan their actions [1, 2].

5. **Assign demographic information to each person.** For each person in a neighborhood i, we assign a race, gender, and age. At the onset, only the age types are not equally distributed, but we will later modify racial distributions to test our model in circumstances where race is an issue.

Our model of a city as a grid has some important features that bear mentioning. For one, note that all parameters are chosen independently of each other. Also, all parameters (including wealth) are discrete for faster coding and to represent the easily avilable information (it is easier and often more useful to classify someone as "middle class" rather than assigning a real number to his or her wealth). Certain parameters are specific to the individual (e.g. race), while some are specific to a neighborhood (e.g. wealth).

8.2.3 Calculating Probabilities and Weights on the Urban Grid

Suppose that we have the crime vectors $\mathfrak{C}_1, \ldots, \mathfrak{C}_N$. We need at least $n = 5$ crime incidents in a serial case, which is a threshold for useful geographic profiling [4]. We assume that every neighborhood on our grid has either zero population data or an appreciable population size, which allows us to use *densities* rather than absolute numbers (e.g. using $\rho_i^{(R)}$ rather than the total number of residents of race R_i) to draw conclusions.

In terms of criminal characteristics, we suppose that the criminal is attacking strangers, consistent with many criminology studies [1]. Serial criminals who attack known associates are much less prone to victim selection patterns, and hence the sporadic nature of their attacks are not suitable for geographic profiling. We assume that the criminal is acting on his own, and his actions are apolitical and not associated with any organization (legal or otherwise). Further, we assume that parameters in the psychological map are independent of one another as viewed by the criminal. Although it is also reasonable to suppose that serial criminals are generally aware of their surroundings, or at least have some working knowledge of the makeup of the nearby environment, our model nonetheless takes the possibility of random action seriously by subtracting the consistent characteristics of victims from the overall relevant population characteristic averages to account for the "state of readiness" to commit a crime [1].

8.2.3.1 Calculation of the Psychological Danger Map

Our five psychological characteristics are race (R), gender (G), age (A), average wealth of neighborhood (ω), and average safety of neighborhood (σ). We first calculate the percentage of the *total* population that has a certain property. Hence, for race A, for example, we have

$$P(R = R_A) = \frac{\sum_{x,y} \rho_A^{(R)}(x,y)P(x,y)}{\sum_{x,y} P(x,y)} \tag{8.6}$$

This is precisely the percentage of those belonging to the group $R = R_A$ across our entire grid. After these probabilities have been calculated, we then compare them to the crime vectors. Given that there are $N_{R=R_A}$ crime vectors $\mathfrak{C}_i$ such that $\mathfrak{C}_i.R = R_A$ out of a total of N crime vectors, we define

$$\pi_{(R=R_A)} = \frac{N_{R=R_A}}{N} \tag{8.7}$$

as the probability that a randomly chosen crime vector will have this characteristic. Then, we consider the value

$$\mathcal{M}_R = \max\{|P(R = R_j) - \pi_{(R=R_j)}|\} \tag{8.8}$$

where j is any race (in our case, $j = A, B, C,$ or D). We do this for every psychological parameter (R, G, A, ω, σ), and consider the set

$$\{\mathcal{M}_R, \mathcal{M}_G, \mathcal{M}_A, \mathcal{M}_\omega, \mathcal{M}_\sigma\} \tag{8.9}$$

These values $\mathcal{M}_i$ we define as the criminal selectivity of i, and are an indication of how randomly the criminal is acting with respect to a certain parameters. This number is between 0 and 1, and measures the difference in how the criminal acts and how he should act given that he does not select for victims based on characteristic K. If the criminal does not select for a particular characteristic, then we expect that the percentage of victims he has of each type for characteristic K will be equal to the percentage of the total population that falls into that type.

The two psychological parameters corresponding to the two largest of the $\mathcal{M}$ values are our key psychological parameters. We will ignore the psychological parameters that are not key psychological parameters; this is because it is more reasonable that a serial criminal will be psychologically consistent in one or perhaps two parameters rather than all five. Suppose K and K' are our two key psychological parameters. We then can calculate both the weight ϕ_p and the value of the psychological danger map δ_p at each neighborhood.

We have

$$\delta_p(a,b) = \frac{1}{2}\left(\sum_i \pi_{(K=K_i)}\rho_i^{(K)}(a,b) + \sum_j \pi_{(K'=K_j)}\rho_j^{(K')}(a,b)\right) \tag{8.10}$$

where the first sum is summed over all possible values of K and the second summed over all possible values of K'. If $K = \sigma$ or $K = \omega$ (i.e. if K is a parameter that characterizes a neighborhood rather than an individual), then for a neighborhood (a,b) where $K(a,b) = K_m$, $\rho_i^{(K)}(a,b) = 0$ for $i \neq m$ and $\rho_m^{(K)}(a,b) = 1$. If, for example, our culprit were to attack only those of race R_A and age A_Y, and if race and age were our key psychological parameters for that criminal, then any neighborhood (a,b) with only individuals belonging to race R_A and age group A_Y will have $\delta_p(a,b) = 1$. Note that $\delta_p \in [0,1]$. Similarly, we define

$$\phi_p = \frac{1}{2}(\mathcal{M}_K + \mathcal{M}_{K'}) \tag{8.11}$$

The psychological weight is thus just defined as the average of the two largest values for criminal selectivityThis definition makes intuitive sense and reduces noise; the values $\mathcal{M}$ describe how much a criminal is selecting for a certain victim characteristic, and hence a very psychologically selective criminal will have by this definition a high value for ϕ_p.

8.2.3.2 Calculation of the Spatial Danger Map and Total Danger Map

For every point $p = (a, b)$ on our grid, we can calculate

$$\langle D(p, C_i) \rangle \tag{8.12}$$

the average distance between p and the crime coordinates C_i. The point p_c such that this average is at a minimum is known as the central point. If there is more than one possible value for p_c, we randomly select one.

We now calculate the values of the spatial danger map. First, we assign a value to every point p on our grid based on its proximity to the crimes, weighted towards the later crimes (consistent with the idea that serial criminals tend to spread out as they commit more crimes, and hence a crime committed later would tend to be closer to other later crimes [15]). Namely,

$$\xi(p) = \sum_{i=1}^{N} (1.1)^i D(p, C_i) \tag{8.13}$$

The 1.1 is our weight factor[3]. We normalize this quantity and form

$$\hat{\xi}(p) = \frac{\sum_{i=1}^{N} (1.1)^i D(p, C_i)}{\max\{\xi\}} \tag{8.14}$$

where by $\max\{\xi\}$ we mean the maximum value of ξ for all points in our grid. Hence, we now have a value between 0 and 1 to describe the proximity of any point to our crime scenes, with weight towards the later crimes.

Next, given all the values $D(p_c, C_i)$ for every C_i, we can define the radius $\mathcal{R}$ of crimes as the average of these values, and the radial variance $\mathcal{V}$ of crimes as the variance of this set of values. Then, for any point $p = (a, b)$, we have that

$$\delta_s(p) = \delta(a, b) = \frac{1}{2}\left(\text{Exp}\left[-\frac{(D(p, p_c) - \mathcal{R})^2}{\sqrt{2\mathcal{V}}}\right] + \hat{\xi}(p)\right) \tag{8.15}$$

The first term is a Gaussian with variance $\mathcal{V}$ and mean $\mathcal{R}$. Hence, the first term produces a value between 0 and 1 describing how much the $D(p, p_c)$ deviates from $\mathcal{R}$. This contribution follows the theory that many criminals operate out of a "home base" [9], and hence their crime scenes tend to orient themselves to be approximately the same distance from this home base[4]. The second term describes how far away a point is from the crime scenes in general, and follows the theory that crime scenes tend to radiate outwards [14], which means that points closer to later crimes are more likely to be chosen as subsequent crime sites. With these two contributions, we can define $\delta_s(p)$, a value between 0 and 1, and the psychological danger of a point.

What remains is to calculate the weight ϕ_s. Given a $k \times \ell$ grid,

$$\phi_s = \text{Exp}\left[-\beta \frac{\mathcal{V}}{\sqrt{k\ell}}\right] \tag{8.16}$$

The parameter $\beta > 0$ depends on qualities of the city. The larger the value for β, the more sensitive to variance the weight ϕ_s becomes. Hence, we would expect β to be larger for locales that are more crowded or for larger grids in general, where the criminal has a larger activity region to function in. For our 10×10 grid, we use $\beta = 5$, and in this case we have

$$\phi_s = \text{Exp}\left[-\frac{\mathcal{V}}{2}\right] \tag{8.17}$$

[3]The weight factor 1.1 is large enough to substantially weight ξ towards the later crimes, but not large enough to completely wash out data from the older crimes.

[4]Note that in this case, distance doesn't mean that we would geometrically expect a circle, since there exists a line of fast transportation down $x = 3$.

and hence a variance of 2 will result in the weight becoming $1/e$.

We can now form the total danger map as

$$\Delta(a,b) = \psi(\phi_s)\delta_s(a,b) + \psi(\phi_p)\delta_p(a,b) \tag{8.18}$$

where

$$\psi(x) = \frac{1+e^4}{e^4-1}\left(\frac{1}{1+e^{-2(k(x-\frac{1}{2}))}} - \frac{1}{1+e^4}\right) \tag{8.19}$$

which assigns to every (a,b) a value between 0 and 2. ψ is recognizable as a logistics curve which is slightly modified so that $\psi(0) = 0$ and $\psi(1) = 1$ (to 4 significant digits). k is a value that will determine how steep the ψ function is, and for our present model we use the value $k = 4$. We apply the ψ function in order to ensure that significant weight values for ϕ_p, ϕ_s are amplified, while less significant weight values are made negligible. With this model, higher values of ϕ_i are more prevalent in the final sum, and as ϕ_i decreases, the function $\psi(\phi_i)$ begins to drop off rapidly. This allows us to gauge differences between the spatial and psychological models with greater clarity than a purely linear model.

8.3 Results and Case Analysis

As described above, our model takes both spatial and psychological data from the crimes, and based on how statistically significant each set of data is, produces two weights ϕ_s and ϕ_p. Thus, there are four possible general cases, based on whether ϕ_s and ϕ_p are each high or low. In limiting cases where only ϕ_s is high, our model produces a high danger value around the locations of the crimes - these cases are less interesting and we omit a full analysis from this paper. We hence present here results after running our model with two cases - one where the criminal's crimes produce a high ϕ_p and high ϕ_s, and one using real data from the Thomas Sweatt arsonist case.

8.3.1 Test Case: The Cunning Murderer

Mr. X, with vengeance in his heart due to perceived wrongs, decides to murder 30-40 year old (A_M) males (G_M) of Race R_A. He feels comfortable committing crimes only small distances from neighborhood X, at coordinate $(6,6)$. However, he does not commit crimes at X, but rather within a distance of between one and two from that point. δ_s (Fig. 2) is weighted by $\psi(\phi_s) = 0.944$, which is not too far from 1. It is farther from 1 since the crime pattern is not perfectly circularly distributed but shows a highly symmetrical spatial cluster with a radius of 1.3. δ_p (Fig. 3) also has a high overall weight of .751, since the criminal is consistent. We see that it highlights the neighborhoods with similar demographics to the intended targets.

The choice of weights makes Δ (Fig. 4) represent the situation correctly. There is a danger zone about X (with X itself having a lower danger than the surrounding neighborhoods), and there is a rapid drop in danger as we move away from the radius, due to the heavy spatial organization. This is a clear reflection of the "buffer zone effect", as explained in [13]; X is relatively safe, but neighborhoods around the average radius of the crime scenes are very dangerous, with a deformation of symmetry due to weighting towards the later crimes.

Recommendation: Station police officers in the very high danger zones and put those neighborhoods on high alert. Put neighborhood X on medium alert, but do not station a police officer there. Do not issue warning to neighborhoods outside of the immediate hot zone. Since Mr. X is both spatially and psychologically organized, and hence has a reasonably high probability of targeting the specific area without living there, the model should not necessarily be used to make a prediction as to the criminal's residence, *despite the fact that the current models would put it at* X. It is possible to implement a search effort within the given radius, given that it is not overly expensive to do so.

8.3.2 Washington DC: Thomas Sweatt Arson Case

Now we apply our model to a real-life case. From 2002 to 2004, Thomas Sweatt set fire to 46 apartments and cars in Washington D.C. First, we split our city into four quadrants, for

which we use available data on race distribution in Washington D.C. to give each quadrant the appropriate distribution in expectation. We assume that gender is uniformly distributed, and that the age distribution from a 2000 census is the same for all four quadrants. Then, we manually overlay a 10×10 grid onto the city map and mark where the crimes were committed. We take careful note of the Washington Post article on the subject, where there is a report that Thomas "told investigators that he chose his targets at random" and that "he was acting largely on impulse while scoping targets in a car" [17]. This suggests that Thomas was spatially organized (insofar as he did not in general get very far from his place of residence before setting a fire) and not psychologically organized, which we see from the random nature of δ_p (Fig. 5), while δ_s (Fig. 6) shows peaks centered at $(7, 4)$.

Since Washington D.C. is far from homogenous [19], the crimes seem to have some consistency in them, which we see in δ_p. However, these values are close to city averages, so the weight of the psychological danger map is only .200, less than half the weight of the spatial danger map (.924). Putting the two together, we obtain a total danger map (Fig. 7) that is very consistent with Thomas' actual future crimes.

Recommendation: Based on these ten crimes, we recommend heavily increased police presence in the immediate hot zone (within a distance of roughly two neighborhoods of $(7, 4)$). The heavy spatial organization as compared to the weak psychological organization suggests a residence search starting at a radius of .9 from the center point. A general community advisory is not suggested outside of the hot zone, since we see that the topography of the danger map is almost completely flat compared to the peak.

Comments on this Case: Considering the rest of the available crime data, we see that the model works incredibly well. The above recommendations would have helped prevent the next two crimes, would have missed crime 13, 14, 16[5], and would have helped with crimes through 22, after which he started radiating outwards. However, we likewise note that once he started radiating out, the heavy time weight of the later crimes compared to the early crimes would have made the radius calculation larger. However, Mr. Sweatt was actually located at $(6, 3)$ - right within the first or second line of search. So, our geographical profiling model would have helped prevent the next few crimes that Thomas wanted to commit, and would ensure that he was caught at his place of residence before he could set the rest of the city on fire!

8.4 Conclusion

As seen in various test cases and in the above, our model generates useful predictions for crimes that are not completely random. While we do not recommend that our model replace good intuition and solid detective work, we have shown that it produces results that complements professional intuition in serial crime cases and helps direct search efforts and warn the inhabitants of neighborhoods in imminent danger from a serial criminal without causing undue widespread panic. The 2005 case of the Washington D.C. arsonist Thomas Sweatt case is a perfect proof of concept. Using mild demographic assumptions informed by recent census data [19] and a manual grid-overlay[6], our model not only predicted the danger area of most of the subsequent offenses (given only the first 10 data points), but also accurately put the residence of Mr. Sweatt on the correct radius. In essence, the search space was limited to about five neighborhoods, which would have facilitated Mr. Sweatt's arrest long before the rest of the string of arsons could unfold.

This outlines the general procedure for using the model and its advantages. A map of the city, complete with necessary parameters, would be put on a grid and synchronized with police information systems to automatically update. For a string of serial crimes, victim information and location would be put into the model, and after five or more data points[7] were entered, the model

[5]Located at $(6, 6)$,$(4, 5)$, and $(6, 2)$, respectively.

[6]Of course, accuracy would be greatly improved and the results would be even clearer in real life due to the data and technology available to law enforcement that allows near-instantaneous localizations and data input on grids.

[7]Fewer are possible, of course, but then the predictions will likely be rather inaccurate due to lack of data

would generate the recommendations similar to those outlined here. Any recommendation would be used to either confirm or modify existing police efforts and community warnings, which would make the city safer.

The dual-layered design of the model provides obvious advantages over the simple existing models. The calculations of ϕ_s and ϕ_p based on the crime data would allow one to automatically and intelligently place criminals on the gradient between selecting for victim location and selecting for "victim type" (e.g. race, gender). The ability to make this distinction allows our profiling system to go one step beyond pre-existing geographic profiling systems and deal with a more diverse pool of serial crimes, saving resources and serving as a tool to pinpoint danger zones even if no prediction about criminal residence is appropriate.

For further research, we suggest that increased accuracy can be achieved if the nature of the crime is considered. For example, we would expect gender to always be a strong factor in serial rape cases, but this factor does not give us much information about the psychological consistency of the culprit; hence, it may be reasonable to omit the contribution from the gender parameter in the calculation of ϕ_p. Likewise, one could take a different account of parameters depending on the area in question. For example, gender may be weighted less in a city where it is too evenly distributed to be much help and would only add noise into the model. Practical spatial considerations, such as accessibility due to roads or railways would impact the likelihood of a criminal being more spatially than psychologically motivated. The psychological equivalent to this, which would make the model more sophisticated, is the use of coupled parameters, i.e. if a culprit exclusively attacks citizens of race R_B and gender G_M, then those neighborhoods with high densities of citizens who are both of race R_B and male can be weighted more heavily as dangerous areas. We hope that further work on this approach will provide a great ally for law enforcement.

References

[1] P. L. Brantingham, and P. J. Brantingham. *Nodes, Paths and Edges: Considerations on the Complexity of Crime and the Physical Environment.* J. Environmental Psychology. (1993) 13, 3–28.

[2] P. L. Brantingham, and P. J. Brantingham. *Notes on the Geometry of Crime.* In P.J. Brantingham, and P.L Brantingham. (eds.), *Environmental Criminology*, (1981) Sage Publications, Beverly Hills, 27–54.

[3] P. L. Brantingham, and P. J. Brantingham. *Residential Burglary and Urban Form.* (1975) Urban Studies. 12, 273–284.

[4] D. Canter. *Confusing Operational Predicaments and Cognitive Explorations: Comments on Rossmo and Snook et al.* Applied Cognitive Psychology. (2005) 19.5, 663–668.

[5] D. Canter, T. Coffey, M. Huntley, and C. Missen. *Predicting Serial Killers' Home Base Using a Decision Support System.* J. Quantitative Criminology. (2000) 16, 457–478.

[6] J. E. Douglas, et al. *Criminal Profiling from Crime Scene Analysis.* Behavioral Sciences and the Law. (1986) 4.4, 401–421.

[7] D. Jamieson. "Why Thomas Sweatt Set Washington on Fire." *AlterNet.* 8 June 2007. Accessed 22 February 2010. http://www.alternet.org/story/53378/

[8] K. Fritzon. *An Examination of the Relationship between Distance Travelled and Motivational Aspects of Arson.* J. Environmental Psychology. (2000) 21, 45–60.

[9] R. N. Kocsis, and H.J. Irwin. *An Analysis of Spatial Patterns in Serial Rape, Arson, and Burglary: The Utility of the Circle Theory of Environmental Range for Psychological Profiling.* Psychiatry, Psychology and Law. (1996) 4.2, 195-206.

and the model would not be as useful.

[10] D. J. Paulsen. *Human Versus Machine: A Comparison of the Accuracy of Geographic Profiling Methods*. J. Investigative Psychology and Offender Profiling. (2006) 3, 77–89.

[11] W. Petherick. *Serial Crime: Theoretical and Practical Issues in Behavioral Profiling*, Academic Press, Massachusetts, 2008.

[12] D. K. Rossmo. *Place, Space, and Police Investigations: Hunting Serial Violent Criminals*. In J.E. Eck , and D.L. Weisburd. (eds.), *Crime and place: Crime prevention studies*, (1995) Vol. 4, Criminal Justice Press, New York, 217–235.

[13] D.K. Rossmo. *Geographic Profiling*, CRC Press, Florida, 2000.

[14] B. Snook, et al. *On the Complexity and Accuracy of Geographical Profiling Strategies*. J. Quantitative Criminology. (2005) 21.1, 1–26.

[15] B. Snook, et al. *Serial Murderers' Spatial Decisions: Factors that Influence Crime Location Choice*. J. Investigative Psychology and Offender Profiling. (2005) 2, 147–164.

[16] J. Warren. et al. *Crime Scene and Distance Correlates of Serial Rape*. J. of Quantitative Criminology. (1998) 14.1, 35–59.

[17] D. W. Wilber, and S. Horwitz. "Targets of Arson Picked Randomly, Investigators Say." *Washington Post*. 29 April 2005. Accessed 21 February 2010. http://www.washingtonpost.com/wp-dyn/content/article/2005/04/28/AR200504280 1403.html

[18] *CBSNews.com* CBS News, Associated Press, Prince George's County. Accessed 22 February 2010. http://www.cbsnews.com/htdocs/fires/arson/map_intro.html

[19] *CensusScope: 2000 Census Data, Charts, Maps, and Rankings* Social Science Data Analysis Network. Accessed 22 February 2010. http://censusscope.org/

FEATURE

9

MY FAVORITE PROBLEM

Sums of Four Squares

Tony Feng[†]
Harvard University '13
Cambridge, MA 02138
tfeng@college.harvard.edu

Lucia Mocz[‡]
Harvard University '13
Cambridge, MA 02138
lmocz@college.harvard.edu

9.1 Introduction

Pick any natural number n. As the 18th-century mathematician Joseph Lagrange proved, you can always write n as a sum of four perfect squares. Not impressed? Try to do the same for three squares — you'll find that it is already impossible at 7. Think a greedy algorithm will work? Just try 23. Still not impressed? Then read on to find out why this seemingly simple theorem makes it as this issue's *My Favorite Problem.*

We begin with a completely elementary proof by descent, a straightforward method exemplifying the apparent simplicity of elementary number theory. Armed with more machinery, we then present solutions via Minkowski's theorem, lattices, and modular forms, demonstrating this theorem's intrinsic connection to modern techniques in algebraic and analytic number theory and geometry. These solutions highlight the beauty of elementary number theory. It is a subject concerned with seemingly simple problems regarding the properties of numbers — integers in particular — with surprising connections and deep consequences for the rest of mathematics.

9.2 First Proof: Method of Descent

In this section, we give a completely elementary proof of the theorem by Fermat's classical method of descent. First, we prove the following lemma.

Lemma 1. *Given a prime p, there exist $a, b \in \mathbb{Z}$ such that $a^2 + b^2 + 1 \equiv 0 \pmod{p}$.*

Proof. Consider the equation

$$a^2 \equiv -1 - b^2 \pmod{p}.$$

There are $(p+1)/2$ possible values for a^2 and $(p+1)/2$ possible values for $-1-b^2$, so they must coincide at some value, by the pigeonhole principle. □

Theorem 2 (Lagrange). *Every natural number can be written as the sum of four perfect squares.*

[†]Tony Feng is a a sophomore mathematics concentrator at Harvard. In terms of math, Tony is mainly in complex analysis and number theory, but I also like philosophy, algorithms, fiction, and cooking desserts.

[‡]Lucia Mocz '13 is a mathematics concentrator at Harvard with a secondary field in music. Her interest in mathematics is twofold: she is allured by the zen-like nature of topology and enticed by the clever problems in number theory. In addition to being the Problems Editor of The HCMR, she was a problem-writer for HMMT and formerly competed in the Siemens and ISEF competitions, where her interest in mathematics developed.

Proof. Note the identity

$$(a^2+b^2+c^2+d^2)(w^2+x^2+y^2+z^2) = (aw+bx+cy+dz)^2 + (ax-bw-cz+dy)^2 + (ay+bz-cw-dx)^2 + (az-by+cx-dw)^2.$$

This shows that the property of being a sum of four squares is closed under multiplication (this identity is motivated by the quaternionic norm). Therefore, it suffices to prove that every prime is a sum of four squares.

Now fix a prime p. By lemma 1, there exist integers x and y such that x^2+y^2+1 is divisible by p. Thus, there exists a number m such that mp may be expressed as the sum of the squares of at most four integers:

$$mp = x_1^2 + x_2^2 + x_3^2 + x_4^2. \qquad (*)$$

We can consider $m < p$ as an additional constraint.[1] We will show that if $m > 1$, then m can be "reduced"; that is to say, we can always find a smaller number $n < m$ such that np can also be expressed as the sum of at most four squares.

We have two cases. Suppose first that m is even. Then $(*)$ is even, and we have three subcases: all four x_k $(k = 1, 2, 3, 4)$ are even, exactly two of the x_k are even, or all four x_k are odd. In each case, the numbers can be paired such that each pair consists of two numbers with the same parity. Without loss of generality, we pair x_1 with x_2 and x_3 with x_4. Then the numbers

$$(x_1+x_2)/2,\ (x_1-x_2)/2,$$

$$(x_3+x_4)/2,\ (x_3-x_4)/2,$$

are integers. Thus

$$\frac{m}{2}\cdot p = \left(\frac{x_1+x_2}{2}\right)^2 + \left(\frac{x_1-x_2}{2}\right)^2 + \left(\frac{x_3+x_4}{2}\right)^2 + \left(\frac{x_3-x_4}{2}\right)^2.$$

That is, $(m/2)p$ can be expressed as the sum of the squares of at most four integers.

Now consider m to be odd. Let y_k $(k = 1, 2, 3, 4)$ be the remainder smallest in absolute value when x_k is divided by m, i.e.:

$$x_k = mq_k + y_k \quad (k = 1, 2, 3, 4)$$

where y_k can be either positive or negative and $|y_k| < m/2$.

We thus have

$$x_k^2 = m^2q_k^2 + 2mq_ky_k + y_k^2 = mQ_k + y_k^2 \quad (k = 1, 2, 3, 4)$$

where $Q_k = mq_k^2 + 2q_ky_k$ is an integer. Therefore,

$$mp = x_1^2 + x_2^2 + x_3^2 + x_4^2 = mq + y_1^2 + y_2^2 + y_3^2 + y_4^2$$

where $q = Q_1 + Q_2 + Q_3 + Q_4$, and

$$y_1^2 + y_2^2 + y_3^2 + y_4^2 = mn$$

where $n = p - q$. We also have $n < m$ since

$$mn = y_1^2 + y_2^2 + y_3^2 + y_4^2 < 4(m/2)^2 = m^2$$

and moreover, $n \neq 0$, or else all the x_k would be divisible by m and thus $(*)$ would necessarily be divisible by m^2, which is impossible since p is prime and $m \neq 1$ and $m < p$.

[1] In the cited lemma, we can select a and b to be both less than $p/2$, that is, so that the sum a^2+b^2+1 is less than p^2 and therefore the quotient m resulting from dividing x^2+y^2+1 by p will be less than p.

We now want to show that np can also be expressed as the sum of not more than four squares. Use the identity cited at the beginning of this solution once more to show that the product $mp \cdot mn = m^2np$ may be expressed as the sum of the squares of four numbers (since mn and mp are each the sum of four squares):

$$m^2np = (x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4)^2 +$$

$$(x_1y_2 - x_2y_1 + x_3y_4 - x_4y_3)^2 + (x_1y_3 - x_3y_1 + x_4y_2 - x_2y_4)^2 + (x_1y_4 - x_4y_1 + x_2y_3 - x_3y_2)^2.$$

We show that both sides of this equality are divisible by m^2. Since $x_k = mq_k + y_k$, each expression in the parentheses on the right side of the equality is divisible by m. Furthermore, in the first set of parentheses, we have that $y_1^2 + y_2^2 + y_3^2 + y_4^2 = mn$, which is divisible by m, and in the remaining three sets, all products of the form y_iy_j cancel. Dividing both sides by m^2, we get

$$np = z_1^2 + z_2^2 + z_3^2 + z_4^2,$$

as desired.

Therefore, if m in $(*)$ is not equal to 1, it can always be decreased, i.e., there will always be a positive $n < m$ such that a similar equality exists. If $n \neq 1$, we can decrease n further. Thus we can always find a sequence of positive integers $m > n > n_1 > n_2 > \ldots$ until we have for natural numbers X_k $(k = 1, 2, 3, 4)$ that

$$p = X_1^2 + X_2^2 + X_3^2 + X_4^2.$$

□

9.3 Second Proof: Geometry of Numbers

The geometry of numbers has a close relation with other fields of mathematics, and an especially interesting one with number theory. Informally, it is the study of convex bodies and integer vectors in n-dimensional space. To understand its connection to this particular problem in number theory, we first develop a little machinery.

A *lattice* in a finite-dimensional euclidean space $\mathbb{R}^n$ is a discrete abelian subgroup that spans $\mathbb{R}^n$; it turns out a lattice in $\mathbb{R}^n$ is necessarily free of rank n.

Now let $L \subset \mathbb{R}^n$ be a lattice. Let Δ be a *fundamental domain* for L, i.e., the smallest area bounded by points in the lattice. If L is the group generated by $\nu_1, \ldots, \nu_n \in \mathbb{R}^n$, then we could, for example, take $\Delta = \{t_1\nu_1 + \cdots + t_n\nu_n, t_i \in [0, 1), \forall i\}$. Then this set has the property that the translates of Δ by elements of L cover the plane and do not overlap. We denote by μ the Lebesgue measure on $\mathbb{R}^n$.

Theorem 3 (Minkowski). *Let $K \subset \mathbb{R}^n$ be a convex region symmetric about the origin with*

$$\mu(K) > 2^n\mu(\Delta).$$

Then K contains a lattice point of L other than the origin.

Proof. We first show that the translates of $\frac{1}{2}K$ by the elements of L are not all disjoint. If we show this, then we will be done. Indeed, we will then have found $p_1, p_2 \in \frac{1}{2}K$ and $\ell_1 \neq \ell_2 \in L$ such that

$$p_1 + \ell_1 = p_2 + \ell_2,$$

which gives

$$\ell_1 - \ell_2 = p_2 - p_1 \in K \tag{9.1}$$

since $p_2 - p_1 \in \frac{1}{2}K - \frac{1}{2}K \subset K$, as K is convex and symmetric about the origin. Then (9.1) is also a nonzero element of L, which is what we want.

So suppose the contrary. Then we have a set of *disjoint* translates $\frac{1}{2}K + \ell$ for $\ell \in L$. As a result, the intersections $\Delta \cap \left(\frac{1}{2}K + \ell\right)$ are disjoint. If we take their measures (with μ denoting the standard Lebesgue measure), we find that

$$\mu(\Delta) \geq \sum_{\ell} \mu\left(\Delta \cap (\frac{1}{2}K + \ell)\right) = \sum_{\ell} \mu\left((\Delta - \ell) \cap \frac{1}{2}K\right) = \mu\left(\frac{1}{2}K\right) = \frac{1}{2^n}\mu(K),$$

a contradiction. Here we are using the fact that the Lebesgue measure μ is translation-invariant and additive for disjoint sets, as well as the fact that the sets $\Delta - \ell, \ell \in L$ cover $\mathbb{R}^2$. □

The proof of Lagrange's theorem is now a straightforward application of Minkowski's theorem and Lemma 1. The key idea to the following solution is a clever selection of our lattice to show that each prime number can be written as a sum of four squares, i.e., $p = a^2 + b^2 + c^2 + d^2$, where (a, b, c, d) is a point in our lattice.

Proof. (Lagrange's theorem). Fix a prime p. By the previous lemma, we may pick a, b such that

$$a^2 + b^2 + 1 \equiv 0 \pmod{p}.$$

Consider the lattice

$$L = \{a_1\nu_1 + a_2\nu_2 + a_3\nu_3 + a_4\nu_4 \mid a_i \in \mathbb{Z}\}$$

where

$$\begin{aligned}
\nu_1 &= (p, 0, 0, 0), \\
\nu_2 &= (0, p, 0, 0), \\
\nu_3 &= (a, b, 1, 0), \\
\nu_4 &= (-b, a, 0, 1).
\end{aligned}$$

The volume of the fundamental parallelogram is

$$\mu(\Delta) = \left|\det \begin{pmatrix} p & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ a & b & 1 & 0 \\ -b & a & 0 & 1 \end{pmatrix}\right| = p^2.$$

Let K be a four-dimensional sphere around the origin with radius $\sqrt{2p}$,

$$K = \{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 \mid x_1^2 + x_2^2 + x_3^2 + x_4^2 < 2p\}.$$

Then $\mu(K) = \frac{1}{2}\pi^2(\sqrt{2p})^4 = 2p^2\pi^2 > 2^4\mu(\Delta)$. By Minkowski's theorem, K contains a nonzero lattice point $(u_1, u_2, u_3, u_4) \in L$. Furthermore, we must have

$$0 < u_1^2 + u_2^2 + u_3^2 + u_4^2 < 2p.$$

Since this is an element of the lattice, we may write

$$(u_1, u_2, u_3, u_4) = a_1\nu_1 + a_2\nu_2 + a_3\nu_3 + a_4\nu_4.$$

Then

$$u_1^2 + u_2^2 + u_3^2 + u_4^2 \equiv a_3^2(a^2 + b^2 + 1) + a_4^2(a^2 + b^2 + 1) \equiv 0 \pmod{p}.$$

So we must actually have $u_1^2 + u_2^2 + u_3^3 + u_4^2 = p$. □

9.4 Third Proof: Modular Forms

Modular forms play a central role in modern number theory. Briefly put, they are a class of very special functions that generalizes the notion of periodic functions. For brevity, we will have to take some facts on faith in this section.

Let Γ be the *modular group* $SL_2(\mathbb{Z})$. For any finite-index subgroup Γ' of Γ, we make the following definition.

Definition 4. A complex function f is called a *modular form* of weight k for Γ' if

- f is holomorphic in the upper half-plane $\mathbb{H}$.
- For $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma'$, we have:

$$f(\gamma \cdot z) = (cz + d)^k f(z)$$

  where γ acts on z by Mobius transformation.

- f is holomorphic at the "cusps" in some sense.

The second condition is the key symmetry property of modular forms. Note that as a special case, it implies $f(z)$ is periodic, and the third condition is a technical point about the resulting Fourier series of f.

Clearly, the constant functions are trivial examples of modular forms. It is not at all obvious that nontrivial modular forms even exist, and indeed this is the point: the symmetry properties of modular forms (of weight k) are so special that they form a *finite-dimensional vector space.* We will not prove this fact here, but it is a standard result in many references (e.g. [1]).

Proof. (Lagrange's theorem). Let $r_4(n) = \#\{(a,b,c,d) \in \mathbb{Z}^4 \mid a^2 + b^2 + c^2 + d^2 = n\}$, i.e. the number of quadruples of integers whose squares add up to n. Let

$$\theta(q) = \sum_{-\infty}^{\infty} q^{n^2} = 1 + 2q + 2q^4 + \dots$$

Then $\theta(q)^4$ is precisely the generating function for $r_4(n)$. Moreover, one can show that

$$\theta(q)^4 \text{ is a modular form of weight 2 and level 4.}$$

Here "level 4" refers to the precise subgroup $\Gamma' \subset \Gamma$ with respect to which $\theta(q)^4$ is a modular form. We also have the following fact:

The space of modular forms of weight 2 and level 4 is a 2-dimensional vector space.

It turns out that

$$\sum_{n \geq 0} \left(8 \sum_{d|n} d - 32 \sum_{4d|n} d \right) q^n$$

is also a modular form of weight 2 and level 4. Therefore, we can verify that the two functions are actually equal by checking the first few terms of their Fourier expansions. This gives an explicit formula: $r_4(n) = 8 \sum_{d|n} d - 32 \sum_{4d|n} d$, which is clearly positive. □

9.5 Conclusion

Lagrange's theorem has even further consequences in the evolution of number theory. Additional solutions exist using Hurwitz quaternions, Aubry's lemma, and a combination of Ramanujan's bilateral formula and Jacobi's triple product identity. There even exists a randomized polynomial-time algorithm (running in $O(\log(n)^2)$ time!) for computing a representation $n = a^2+b^2+c^2+d^2$ for a given integer n.

But given such a representation, is it unique? No, for we immediately see that

$$4 = 1^2 + 1^2 + 1^2 + 1^2 = 2^2 + 0^2 + 0^2 + 0^2.$$

We next ask the natural question of just how many ways we can write a number as the sum of four squares. In fact, Jacobi proved a much stronger result of Lagrange's theorem, that the number of ways in which a positive integer can be so written equals 8 times the sum of its divisors that are not multiples of 4 (can you see why Lagrange's theorem follows from this as a direct corollary?). We encourage the enthusiastic reader to find further generalizations of this problem and to explore its wide set of solutions.

References

[1] F. Diamond and J. Shurman. *A First Course in Modular Forms*. Springer, 2005.

[2] S. Lang. *Algebraic Number Theory*. Springer-Verlag, 1994.

[3] A. Lozano-Robledo. Untitled seminar notes on Minkowski's theorem. Available at `http://math.bu.edu/people/alozano/seminar/minkowski.pdf`.

FEATURE

10 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, 2011*. We encourage all submitters to typeset their submissions in LaTeX and submit the source code along with the pdf.

A11 – 1. Let a, b, c be positive real numbers. Prove that:

$$\frac{\sqrt{a^3+b^3}}{a^2+b^2}+\frac{\sqrt{b^3+c^3}}{b^2+c^2}+\frac{\sqrt{c^3+a^3}}{c^2+a^2}\geq\frac{6(ab+bc+ac)}{(a+b+c)\sqrt{(a+b)(b+c)(c+a)}}$$

Proposed by Tuan Le (Fairmont High School, Anaheim, CA)

A11 – 2. Are there any simple groups of order $p(p+1)$, where p is prime?

Proposed by Eric Larson '13.

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.

Editor's Note: The following problem from the previous issue is released again as it received no solutions.

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

11 Solutions

All Things Being Inequalities

S08 – 4. Consider a, b, c three arbitrary positive real numbers. Prove that

$$\sum_{cyc}\sqrt{\frac{b+c}{a}} \geq 2\left(\sum_{cyc}\sqrt{\frac{a}{b+c}}\right)\cdot\sqrt{1+\frac{(a+b)(b+c)(c+a)-8abc}{4\sum_{cyc}a(a+b)(a+c)}}.$$

Proposed by Cosmin Pohoata (Bucharest, Romania).

Solution by Greg Yang '14. Note that if $\sum$ appears without subscript, it will denote a sum over all symmetric variants of the polynomial that follows. For example, $\sum ab = ab+bc+ca$ while $\sum a^2b = a^2b+a^2c+b^2a+b^2c+c^2a+c^2b$. Explicitly, if the monomial that follows has 3 different exponents, then there are 6 terms in the sum; otherwise there are 3 terms in the sum. Similarly for $\prod$ without subscript.

We will replace the $4\sum_{\text{cyc}}a(a+b)(a+c)$ with $3\sum_{\text{cyc}}a(a+b)(a+c)$ to prove a stronger version of the problem statement. Manipulate the inequality into the following form:

$$\frac{\sum_{\text{cyc}}\sqrt{\frac{a+b}{c}}}{2\sum a} \geq \frac{\left(\sum_{\text{cyc}}\sqrt{\frac{a}{b+c}}\right)}{\sum a}\sqrt{1+\frac{(a+b)(b+c)(c+a)-8abc}{3\sum_{\text{cyc}}a(a+b)(a+c)}}.$$

Let $f(x) = x^{-1/2}$. f is convex for $x > 0$. By weighted Jensen, the LHS is

$$\frac{\sum_{\text{cyc}}(b+a)f((a+b)c)}{\sum a+b} \geq f\left(\frac{\sum_{\text{cyc}}(b+a)^2c}{2\sum a}\right) = \sqrt{\frac{2\sum a}{\sum a^2b+6\prod a}}.$$

On the other hand, let $g(x) = x^{1/2}$. g is concave for $x > 0$. Again by weighted Jensen,

$$\begin{aligned}\frac{\left(\sum_{\text{cyc}}\sqrt{\frac{a}{b+c}}\right)}{\sum a} &= \frac{\sum_{\text{cyc}}ag\left(\frac{1}{(c+b)a}\right)}{\sum a}\\ &\leq g\left(\frac{\sum_{\text{cyc}}\frac{a}{(c+b)a}}{\sum a}\right) = \sqrt{\frac{\sum_{\text{cyc}}\frac{1}{c+b}}{\sum a}} = \sqrt{\frac{\sum_{\text{cyc}}(a+b)(b+c)}{(\sum a)(\prod(a+b))}}\\ &= \sqrt{\frac{\sum a^2+3\sum ab}{(\sum a)(\prod(a+b))}}.\end{aligned}$$

Hence it suffices to prove

$$\sqrt{\frac{2\sum a}{\sum a^2b+6\prod a}} \geq \sqrt{\frac{\sum a^2+3\sum ab}{(\sum a)(\prod(a+b))}}\sqrt{1+\frac{(a+b)(b+c)(c+a)-8abc}{3\sum_{\text{cyc}}a(a+b)(a+c)}}.$$

Since

$$\begin{aligned}3\sum_{\text{cyc}} a(a+b)(a+c)+\prod(a+b)-8abc &= 3\left(\sum a^3+\sum a^2b+3abc\right)\\ &\quad+\left(\sum a^2b+2abc\right)-8abc\\ &= 3\left(\sum a^3+\sum a^2b\right)+\left(\sum a^2b+3abc\right)\\ &= \left(\sum a\right)\left(3\sum a^2+\sum ab\right),\end{aligned}$$

(note that factorability here is the reason we replaced the 4 with the 3), we have

$$\begin{aligned}&\left(2\sum a\right)\left(\prod(a+b)\right)\left(3\sum a(a+b)(a+c)\right)\\ &\qquad\geq\left(\sum a^2b+6\prod a\right)\left(\sum a^2+3\sum ab\right)\left(3\sum a^2+\sum ab\right).\end{aligned}$$

Now we painfully expand. The easiest way is to note that the expansion is a 7-degree polynomial and then count the coefficient of each 7-degree monomial on each side. After expansion we get

$$3\sum a^6b+5\sum a^5b^2+5\sum a^4b^3-2\sum a^5bc-14\sum a^4b^2c-2\sum a^3b^3c+2\sum a^2b^2c^3\geq 0.$$

But we have by AM-GM

$$\frac{1}{3}a^3b^3c+\frac{2}{3}a^6c\geq a^5bc \quad\Rightarrow\quad \frac{2}{3}\sum a^3b^3c+\frac{2}{3}\sum a^6c\geq 2\sum a^5bc \tag{11.1}$$

$$\frac{1}{2}a^2b^2c^3+\frac{1}{2}c^5b^2\geq c^4b^2a \quad\Rightarrow\quad \sum a^2b^2c^3+\frac{1}{2}\sum c^5b^2\geq\sum c^4b^2a \tag{11.2}$$

$$\frac{2}{3}a^3b^3c+\frac{1}{3}a^6c\geq a^4b^2c \quad\Rightarrow\quad \frac{4}{3}\sum a^3b^3c+\frac{1}{3}\sum a^6c\geq\sum a^4b^2c. \tag{11.3}$$

Using one of (11.1) and (11.3) and two of (11.2), the inequality simplifies to

$$2\sum a^6b+4\sum a^5b^2+5\sum a^4b^3-\sum 11a^4b^2c\geq 0.$$

Now this inequality is true by Muirhead as $(6,1,0),(5,2,0),(4,3,0)$ all majorize $(4,2,1)$. □

Cruel Mistress Induction

F08 – 1. Let p, q be two positive integers, and let n be integers such that $n\geq p+q$. Prove that the following identity holds:

$$\sum_{i=0}^{p}\binom{p}{i}\binom{q}{p-i}\binom{n+i}{p+q}=\sum_{i=0}^{p}\binom{p}{i}\binom{n}{i}\binom{n-i}{q}.$$

Proposed by Cosmin Pohoata (Bucharest, Romania).

Solution by Arnav Tripathy '11. The right-hand side is

$$\sum_{i=0}^{p}\binom{p}{p-i}\binom{n}{q}\binom{n-q}{i}=\binom{n}{q}\binom{n-q+p}{p}.$$

where the last step follows from noting that after factoring out a $\binom{n}{q}$ we are, in effect, just counting the number of ways to choose p people from $n-q$ males and p females in two different ways.

Hence, the problem follows by substituting $r = n + p$ (and replacing i with $p - i$) in the following claim: for any $r \geq p + q$,

$$\sum_{i=0}^{\infty} \binom{p}{i}\binom{q}{i}\binom{r-i}{p+q} = \binom{r-p}{q}\binom{r-q}{p}$$

which is in turn the special case $d = 0$ of the following more general claim: for any $0 \leq d \leq p, r \geq p + q - d$, we have

$$\sum_{i=0}^{p} \binom{p}{i}\binom{q}{i}\binom{r-i}{p+q-d} = \sum_{j=0}^{d} \binom{d}{j}\binom{r+j-p}{q-d+j}\binom{r-q}{p-j}$$

Indeed, denote the left- and right-hand sides above by $f(p, q, r, d)$ and $g(p, q, r, d)$, respectively. We prove that $f(p, q, r, d) = g(p, q, r, d)$ by simultaneous induction on r and reverse induction on d. In other words, we first note that the identity holds for r as small as possible, i.e. $r = p + q - d$, where $f(p, q, r, d) = g(p, q, r, d) = 1$, and for d as large as possible, i.e. for $d = p$, where we have

$$f(p, q, r, d) = \sum_{i=0}^{p} \binom{p}{i}\binom{q}{i}\binom{r-i}{q} = \sum_{i=0}^{p} \binom{p}{i}\binom{r-i}{q-i}\binom{r-q}{i} = g(p, q, r, d)$$

as desired.

We finish by noting that

$$f(p, q, r, d) = f(p, q, r-1, d) + f(p, q, r-1, d+1)$$

and

$$g(p, q, r, d) = g(p, q, r-1, d) + g(p, q, r-1, d+1)$$

so that the induction step follows. Indeed, using $\Delta h(r)$ to denote $h(r+1) - h(r)$, we have

$$\begin{aligned}
\Delta f(p, q, r, d) &= \sum_{i=0}^{p} \tbinom{p}{i}\tbinom{q}{i}\Delta\tbinom{r-i}{p+q-d} = \sum_{i=0}^{p} \tbinom{p}{i}\tbinom{q}{i}\tbinom{r-i}{p+q-d-1} \\
&= f(p, q, r, d+1) \\
\Delta g(p, q, r, d) &= \sum_{j=0}^{d} \tbinom{d}{j}\left(\Delta\tbinom{r+d-j-p}{q-j}\tbinom{r-q}{p-d+j} + \tbinom{r+d-j-p+1}{q-j}\Delta\tbinom{r-q}{p-d+j}\right) \\
&= \sum_{j=0}^{d} \tbinom{d}{j}\left(\tbinom{r+d-j-p}{q-j-1}\tbinom{r-q}{p-d+j} + \tbinom{r+d-j-p+1}{q-j}\tbinom{r-q}{p-d+j-1}\right) \\
&= g(p, q, r, d+1)
\end{aligned}$$

as claimed. □

Also solved by Greg Yang '14

Editor's Note: The following problem required a correction from the previous issue. A solution by the editor is presented for the corrected problem.

Fun with Fermat

F08 – 2. Let p be an odd prime. For every positive integer n, let

$$A(n) = 1^n + 2^n + \cdots + (p-2)^n \quad \text{and} \quad B(n) = 1^n + (p-1)^n.$$

Let $\{a_i\}_{i=1}^{\infty}$ be the sequence defined by $a_1 = 2, a_2 = p^2 + 2$ and

$$\begin{cases} a_{n+2} = A(n)a_{n+1} + B(n)a_n & \text{if } p-1 \nmid n, \\ a_{n+2} = [A(n) + B(n)]a_{n+1} + a_n & \text{if } p-1 \mid n. \end{cases}$$

Prove that no a_n is equal to the product of any $p-1$ terms of the sequence $\{a_i\}_{i=1}^{\infty}$.

Proposed by Daniel Campos Salas (Costa Rica).

Solution by Lucia Mocz '13. In this proof, we rely heavily on the result of the following theorem:

Fermat's Little Theorem. If p is a prime number, then for any $a \in \mathbb{Z}$, $a^p - a$ is evenly divisible by p. In modular notation, this can be stated as $a^p \equiv a \mod p$. A variant to this statement is that if $a \in \mathbb{Z}$ is coprime to p then $a^{p-1} - 1$ is evenly divisible by p. Again, in modular notation, we write this as $a^{p-1} \equiv 1 \mod p$.

By this theorem, if we show that each term in the sequence is congruent to 2 modulo p, then the product of any $p-1$ terms is congruent to 1 modulo p. It follows that as the product of any $p-1$ terms in the sequence is not congruent modulo p to any a_n, no a_n can be equal to the product of any $p-1$ terms of the sequence.

We thus proceed by induction on n to show that all terms of the sequence are congruent to 2 modulo p. We already have this is true for $n = 1$ and $n = 2$. Assume it holds for n and $n+1$. We want to show it also holds for $n+2$. Then we have the following two cases:

1. First suppose that $p - 1 \nmid n$. Then we have

$$a_{n+2} = A(n)a_{n+1} + B(n)a_n \equiv 2(A(n) + B(n)) \mod p$$

We want to show that $A(n) + B(n) \equiv 1 \mod p$. From the previous calculation, we know

$$A(n) + B(n) \equiv 1 + (1^n + 2^n + \ldots + (p-2)^n + (p-1)^n) \equiv 1 + x \mod p$$

Note that since $\mathbb{Z}/p\mathbb{Z}$ is a field, then $(\mathbb{Z}/p\mathbb{Z})^\times$ (i.e., its group of units) is cyclic and acts on $\mathbb{Z}/p\mathbb{Z}$. Let $\langle g \rangle = (\mathbb{Z}/p\mathbb{Z})^\times$, where g generates the group. Then:

$$g^n \cdot x \equiv g^n + (2g)^n + \cdots + (g(p-1))^n \equiv x \mod p$$

It follows that $x(g^n - 1) \equiv 0 \mod p$. However, since $p - 1 \nmid n$, we must have $g^n \not\equiv 1 \mod p$, which implies that $x \equiv 0 \mod p$. We are thus left with $A(n) + B(n) \equiv 1 \mod p$ as desired.

2. Now suppose that $p - 1 | n$. Then we have

$$a_{n+2} = (A(n) + B(n))a_{n+1} + a_n \equiv 2(A(n) + B(n)) + 2 \mod p$$

We want to show that $A(n) + B(n) \equiv 0 \mod p$. We have, as in the previous case:

$$A(n) + B(n) \equiv 1 + (1^n + 2^n + \ldots + (p-2)^n + (p-1)^n) \equiv 1 + x \mod p$$

Since $p-1|n$, by Fermat's Little theorem, we have:

$$x \equiv (1^k)^{p-1} + (2^k)^{p-1} + \ldots + ((p-1)^k)^{p-1} \equiv p-1 \mod p$$

where $k = \frac{n}{p-1}$. We are thus left with $A(n) + B(n) \equiv 1 + (p-1) \equiv 0 \mod p$ as desired. □

On a Rolle

F08 – 3. Let $f : [0,1] \to \mathbb{R}$ be a differentiable function with continuous derivative such that

$$\int_0^1 f(x)\,dx = \int_0^1 xf(x)\,dx.$$

Prove that there exists $\xi \in (0,1)$ such that

$$f(\xi) = f'(\xi)\int_0^\xi f(x)\,dx.$$

Proposed by Cezar Lupu (University of Bucharest, Bucharest, Romania).

Solution by Paolo Perfetti (Dipartimento di Matematica, Universita degli di Tor Vergata Roma, Italy). If $f \equiv 0$ there is nothing to prove, so we suppose $f \not\equiv 0$. Define $F(x) = \int_0^x f(y)dy$ and observe that

$$\int_0^1 F(x)dx = \int_0^1 dx \int_0^x f(y)dy = \int_0^1 dyf(y) \in_y^1 dx = \int_0^1 dyf(y)(1-y) = 0$$

We claim the set of points of $[0,1]$, say N, where $F(x) = 0$, is not dense on $[0,1]$. For otherwise we would have $F(x) = 0$ on a dense set in $[0,1]$ and then $F(x) \equiv 0$ by continuity of $F(x)$, but this is forbidden by $f \not\equiv 0$ via the continuity of $f(x)$.

As a consequence, there exists an open interval $(a,b) \subset [0,1]$ where $F(x) \neq 0$. Now define:

$$g : [0,1]\backslash N \to R$$

$$g(x) = f(x) - \ln|F(x)|$$

We affirm that there exists the point $x_1 \in [0,a]$, nearest to $x = a$, such that $\lim_{x\to x_1^+} g(x) = +\infty$ and a point $x_2 \in [b,1]$, nearest to $x = b$, such that $\lim_{x\to x_2^-} g(x) = +\infty$ (this is proved in a moment). By continuity of $g(x)$ in the interval (x_1,x_2), for any $y = y_0$ large enough there are two points ν_1 and ν_2 both in (x_1,x_2) such that $g(\nu_1) = g(\nu_2)$. Rolle's theorem tells us that $g'(\xi) = 0$ at a point $\xi \in (\nu_1,\nu_2)$ yielding

$$f' - \frac{F'(\xi)}{F(\xi)} = 0$$

or

$$f(\xi) = f'(\xi)\int_0^\xi f(x)dx.$$

the existence of x_1. Let $A = \{x \in [0,a] : F(x) = 0\}$. A is nonempty since $F(0) = 0$ and bounded and then there exists $\sup A$ which is attained (is a maximum) since $F(x)$ is continuous. The proof of existence of x_2 is quite similar, but notice that it is essential $\int_0^1 F(x)dx = 0$ because otherwise the set of points $B = \{x \in [b,1] : F(x) = 0\}$ might be empty. □

Also solved by Arnav Tripathy '11

Oh, the Choices You'll Make

F08 – 4. Do there exist functions $f, g : \mathbb{R} \to \mathbb{R}$ such that

- both are periodic, *i.e.* there exist nonzero real a, b such that for all $x \in \mathbb{R}$, $f(x) = f(x+a)$ and $g(x) = g(x+b)$, and
- their sum is equal to the identity, *i.e.* for all $x \in \mathbb{R}$, $f(x) + g(x) = x$?

Proposed by Robert Obryk (August Witkowski High School, Krakow, Poland).

Solution by Oliver Knill (Harvard University). We can assume $a = 1$ by rescaling the axes. We can assume b irrational because for rational $b = p/q$ the condition $f(x) + g(x) = f(x + qb) + g(x + qb) = x + qb$ collides with $f(x) + g(x) = x$. The condition $f(x) + g(x) = x$ means

$$f(x + kb + l) + g(x + kb + l) = x + kb + l$$

for all x and all integers k, l. By the $a = 1$ periodicity of f and the b periodicity of g, this is

$$f(x + kb) + g(x + l) = x + kb + l. \tag{11.4}$$

Define an equivalence relation on the reals by

$$x \sim y \Leftrightarrow x + kb + l = y \text{ for some integers } k, l.$$

Let X be the quotient $X = \mathbb{R}/\sim$. We can chose a member $x(s)$ in each equivalence class of X by the axiom of choice. Define $f(x(s)) = x(s)$ and $g(x(s)) = 0$. Now the values of f and g on each equivalence class is determined with

$$f(x(s) + kb) = f(x(s) + kb + l) = x(s) + kb$$
$$g(x(s) + l) = g(x(s) + kb + l) = l$$

This defines the values of f and g for every x because every real x can be written as $x = x(s) + kb + l$, with $s \in X, k, l \in \mathbb{Z}$. Adding these two equations gives (11.4). Note that the functions f, g are given in a nonconstructive way and non Lebesgue measurable but they do exist if one accepts the axiom of choice. □

Also solved by the Missouri State University Problem Solving Group and Arnav Tripathy '11

Three's a Charm

F08 – 5. Let ABC be an arbitrary triangle and let I be the incenter of ABC. Let D, E, F be the points on lines $\overline{BC}, \overline{CA}, \overline{AB}$ respectively such that $\angle BID = \angle CIE = \angle AIF = 90°$, and define the following measurements: r_a, r_b, r_c are the exradii of the triangle ABC, Δ' is the area of DEF, and Δ is the area of ABC. Prove that

$$\frac{\Delta'}{\Delta} = \frac{4r(r_a + r_b + r_c)}{(a+b+c)^2}.$$

Proposed by Mehmet Şahin (Ankara, Turkey).

Solution by Kee-Wai Lau (Hong Kong, China). Let $a = BC$, $b = CA$, $c = AB$. Denote by r and s, respectively, the inradius and semiperimeter of triangle ABC. It is well-known that $r = \frac{\Delta}{s}$. We have

$$AF = AI \sec(A/2) = r \csc(A/2) \sec(A/2) = \frac{2r}{\sin A} = 2\left(\frac{\Delta}{s}\right)\left(\frac{bc}{2\Delta}\right) = \frac{bc}{s}$$

and similarly $CE = ab/s$.

Hence $AE = AC - EC = b - ab/s = b(s-a)/s$ and the area of triangle AFE is:

$$\frac{(AF)(AE)\sin A}{2} = \frac{1}{2}\left(\frac{bc}{s}\right)\left(\frac{b(s-a)}{s}\right)\left(\frac{2\Delta}{bc}\right) = \frac{b(s-a)\Delta}{s^2}.$$

Similarly, the area of triangle BDF and CED are $c(s-b)\Delta/s^2$ and $a(s-c)\Delta/s^2$. Thus

$$\Delta' = \Delta = \frac{b(s-a)\Delta + c(s-b)\Delta + a(s-c)\Delta}{s^2} = \frac{2(ab+bc+ca) - (a^2+b^2+c^2)}{(a+b+c)^2}\Delta.$$

It remains therefore to show that $4r(r_a + r_b + r_c) = 2(ab+bc+ca) - (a^2+b^2+c^2)$. Using the well-known results:

$$r_a = \frac{\Delta}{s-a}, r_b = \frac{\Delta}{s-b}, r_c = \frac{\Delta}{s-c}$$

and Heron's formula for Δ, we see that:

$$\begin{aligned}
4r(r_a + r_b + r_c) &= \frac{4\Delta}{s}\left(\frac{\Delta}{s-a} + \frac{\Delta}{s-b} + \frac{\Delta}{s-c}\right) \\
&= \frac{4\Delta^2\left((s-b)(s-c) + (s-c)(s-a) + (s-a)(s-b)\right)}{s(s-a)(s-b)(s-c)} \\
&= (a-b+c)(a+b-c) + (a+b-c)(-a+b+c) + (-a+b+c)(a-b+c) \\
&= 2(ab+bc+ca) - (a^2+b^2+c^2)
\end{aligned}$$

as desired. □

Also solved by Arnav Tripathy '11

FEATURE

12

ENDPAPER

The Three-Legged Theorem

Michael J. Hopkins[†]
Harvard University
Cambridge, MA 02138
mjh@math.harvard.edu

The summer after my senior year in high school, I had a job driving a truck for a small company in Nebraska. I'd start every day in Fremont, pull orders, load a truck and drive it to Lincoln. There I would unload the truck, shoot the crap with the guys in the warehouse, re-load and head to Council Bluffs. From Council Bluffs, I drove back to Fremont. It was a pretty good job if you like driving around in the country and shooting the crap with guys in warehouses. The only problem was that I had trouble staying awake behind the wheel.

I tried several strategies. My first was to yell "hey!" and point every time I saw a bale of hay. My father did this all the time on family trips and I thought maybe I was at last old enough to do it myself. Evidently I wasn't.

My next idea was to insult the cows. There are a lot of them scattered around the Nebraska countryside and mostly they just kind of stand there. I thought it would be really cool to have a cow to go, "wtf?!" This worked pretty well. I could get really worked up yelling insults at cows and it definitely kept me awake. I pulled into Lincoln the first time I tried it, pumped up and in a great mood. In the middle of shooting the crap, one of the warehouse guys asked me why I was so excited and so hoarse. I stalled a bit before coming up with, "big night."

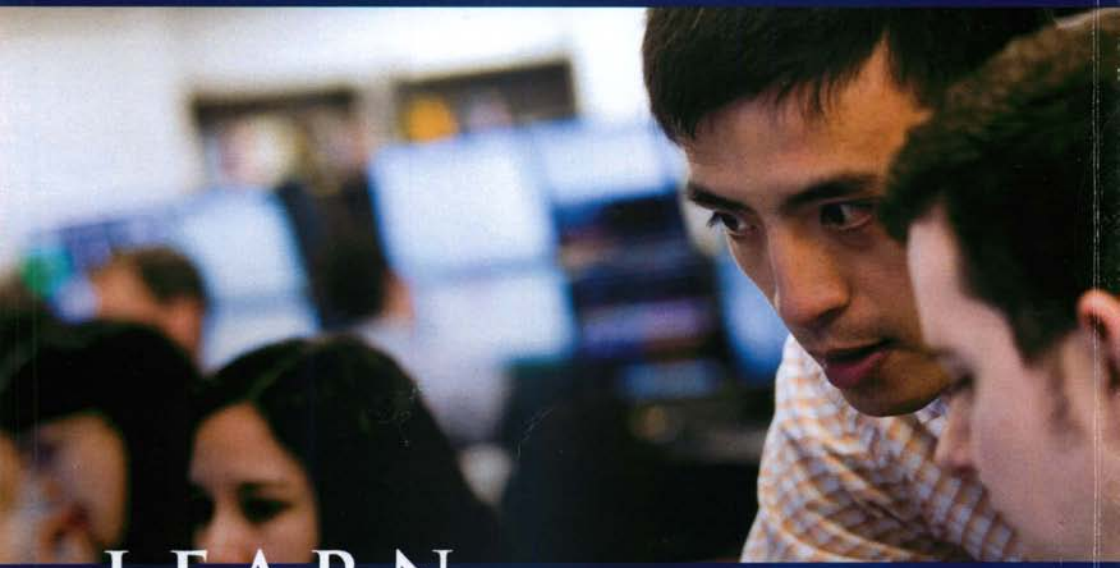
This went on for a couple of weeks during which I got a lot of respect from the warehouse guys, but no reaction at all from the cows. I started to wonder why. I had two theories. The first was that cows are stupid and that they couldn't hear me from the road anyway. I quickly dismissed that as ridiculous. A more realistic explanation, I reasoned, was that cows are incredibly chill and that they are born nearly enlightened, with a natural understanding that the world is an illusion and life is suffering. They probably felt pity for me. That seemed about right, but it pretty much killed my strategy for staying awake. I felt too guilty. That, and the fact that serenely slipping past fields of bovine buddahs was way too zen to keep me awake.

I needed a new idea and there happened to be one in my lunch bag. It was the textbook for a class in point set topology I had taken at UNO that spring. I decided to road test my understanding, starting from page one. Before each trip I'd read the statement of a theorem and then try and prove it on the road. For a while I was doing pretty well. Then I got to the Heine-Borel Theorem.

The Heine-Borel theorem is the one characterizing compact subsets of Euclidean space. It's not that the proof is so hard, but you really have to be organized in your mind to come up with it. I wasn't. I got all the way to Lincoln without getting anywhere at all. The warehouse guys asked me why I was so distracted. All I could come up with was, "big night." I got back in the truck and drove to Council Bluffs. Still nothing. When I finally got back to Fremont I had to peek. In the end, I drove all three legs of my trip without even coming up with a clue on where to start.

This was my first encounter with a "three-legged theorem." They are something pretty special: little theorems that aren't that hard, but require really clear thinking. You meet them from time to time, and if you're lucky, you cannot get anywhere at all trying to prove them. That's when you learn something. Now that you know what to look for, keep your eyes open. They are all over the place.

[†]Michael Hopkins has a PhD from Northwestern University and a D. Phil from Oxford University. He taught at Princeton, the University of Chicago and MIT before coming to Harvard in 2004.



LEARN TRADE TEACH

+ **QUANTITATIVE TRADING AT JANE STREET** WILL CHALLENGE YOUR SKILLS IN A DYNAMIC ENVIRONMENT THAT PRIZES THE DEVELOPMENT OF NEW IDEAS AND TRADING STRATEGIES.

+ **JOIN THE FIRM** FOR THE FASCINATING PROBLEMS, CASUAL ATMOSPHERE, AND HIGH INTELLECTUAL QUALITY. STAY FOR THE CLOSE-KNIT TEAM ENVIRONMENT, GENEROUS COMPENSATION, AND ENDLESS OPPORTUNITIES TO LEARN, TEACH, AND CREATE.

NO FINANCE EXPERIENCE IS NECESSARY

ONLY INTELLECTUAL CURIOSITY AND THE DESIRE TO LEARN.

APPLY ONLINE:

http://janestreet.com/apply/

JANE STREET

WWW.JANESTREET.COM

NEW YORK
LONDON
HONG KONG