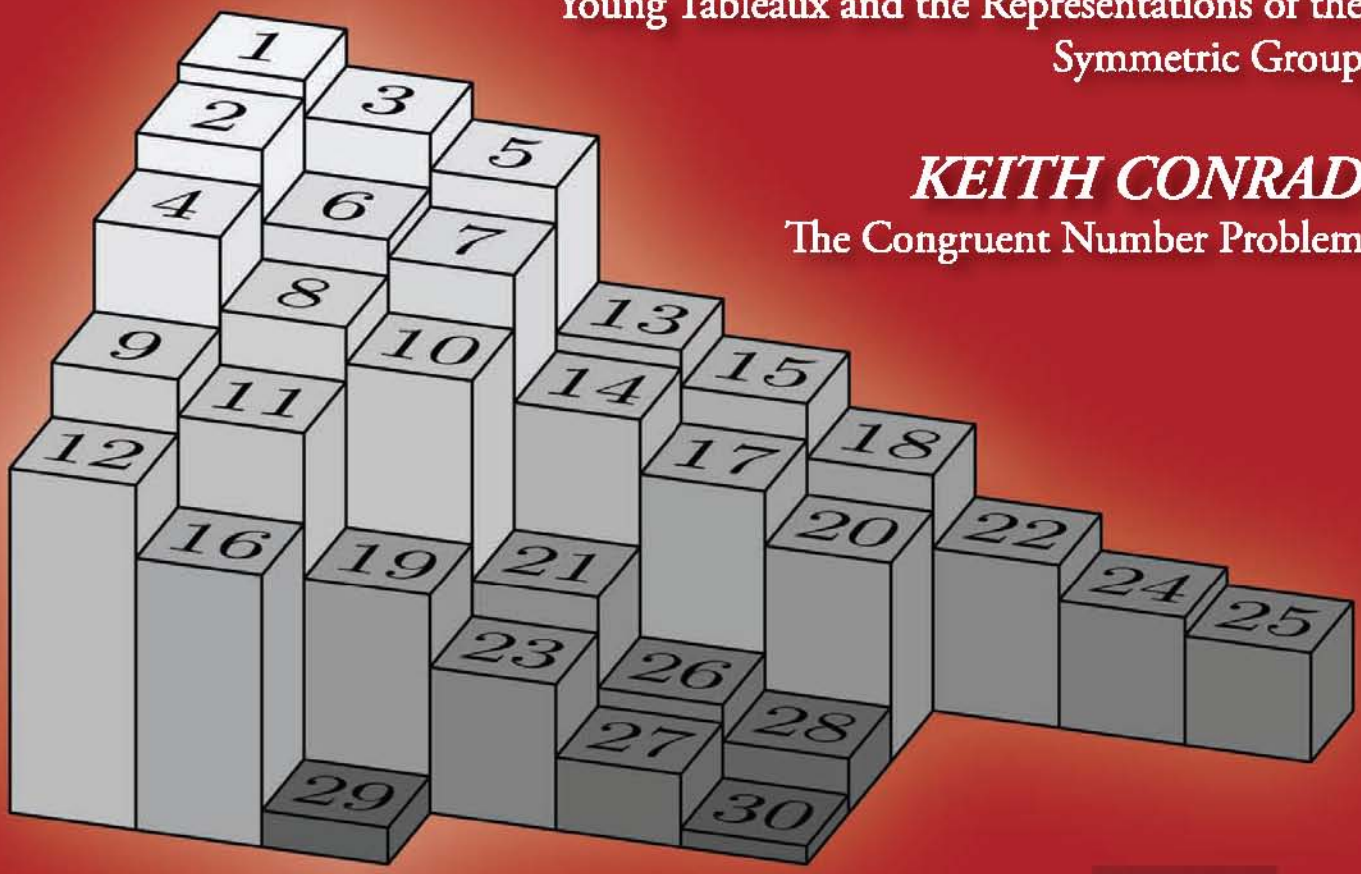
The Harvard College Mathematics Review

Vol. 2, No. 2

Fall 2008

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 Editors-In-Chief Zachary Abel and Ernest E. Fontes 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 Zachary Abel 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 Oluwadara Johnson.

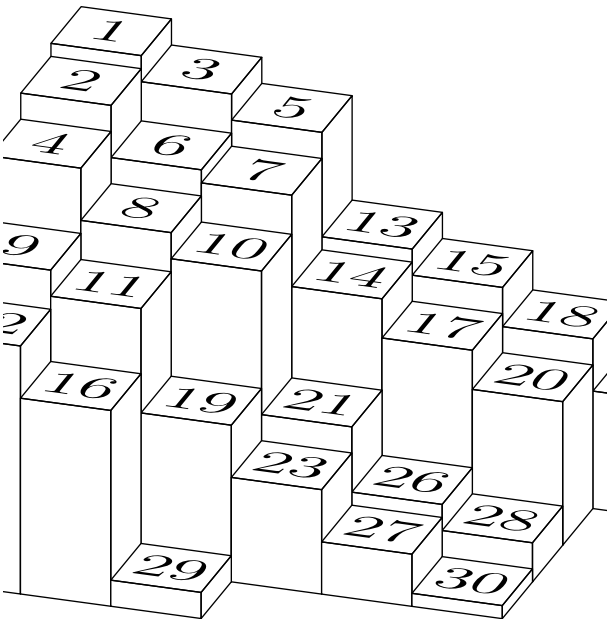
Subscriptions. One-year (two issue) subscriptions are available, at rates of $10.00 for students, $15.00 for other individuals, and $30.00 for institutions. Subscribers should mail checks for the appropriate amount to The HCMR's postal address; confirmation e-mails or queries should be directed to hcmr-subscribe@hcs.harvard.edu.

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

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

Contributors · The Harvard Undergraduate Council · American Mathematical Society · **Patrons** · The Harvard University Mathematics Department

Cover Image. The image on the cover depicts a Young tableaux with towers over each box to illustrate the defining property that values increase (equivalently, heights decrease) in the rightward and downward directions. This issue's article "Young Tableaux and the Representations of the Symmetric Group" by Yufei Zhao (p. 33) studies combinatorial properties and applications of Young tableaux. The image was created in Asymptote™ by Graphic Artist Zachary Abel.



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

-2
Contents

-1
Staff

0 From the Editors

Zachary Abel
Harvard University '10
Cambridge, MA 02138
zabel@fas.harvard.edu

Ernest E. Fontes
Harvard University '10
Cambridge, MA 02138
efontes@fas.harvard.edu

This fall has been marked by the first major transition of leadership for *The Harvard College Mathematics Review* (HCMR). And with this transition, The HCMR has blossomed from an upstart student publication to an established organization. Many of our founding members are now applying to graduate schools and will soon be leaving the journal's front-line operations. Guided by their precedent, we welcome and look forward to the contributions of the many new members that have joined the organization, with whose help we may continue serving as a resource to the undergraduate mathematical community.

By our combined involvement in various roles in The HCMR's staff including Problems Editor and Issue Production Director, we two have been privileged to witness and help the journal grow into its current state and are honored to guide the organizations for this exciting academic year. Our optimism springs from the tireless ingenuity of our contributing student and faculty **authors**, the continued devotion of our **reviewing and editing staffs**, and the indefatigable zeal of our **production directors**. These first four issues are a testament to your support, without which the journal could never have come so far.

We also owe deep gratitude to **Professor Peter Kronheimer** and **Professor Benedict H. Gross '71** for their advice and guidance, **Professor Clifford H. Taubes** for continued encouragement, and the rest of The HCMR's **advisors** and **sponsors**, whose profound contributions have been a foundation for the journal's success. Our executive board owes much to the administrative assistance of **Dean Paul J. McLoughlin II**, **David R. Friedrich** and the rest of the staff at the Student Organization Center at Hilles, and to the unceasing, generous support of the **Harvard Mathematics Department**.

Finally, the two of us would like to express our gratitude to Editor Emeritus **Scott D. Kominers '09**, who continues to offer invaluable expertise and guidance to The HCMR even though he has stepped down from his position as Editor-In-Chief. Scott was an inspirational leader and founder of The HCMR, and he will always be remembered fondly by our staff, authors, and countless readers who have been touched by his work in the journal. Thank you, Scott, and farewell.

Zachary Abel '10 and Ernest E. Fontes '10
Editors-In-Chief, The HCMR

STUDENT ARTICLE

1 Error-Correcting Codes and Sphere Packings

François Greer[†]
Harvard University '11
Cambridge, MA 02138
fgreer@fas.harvard.edu

Xiaoqi Zhu[‡]
Harvard University '11
Cambridge, MA 02138
xzhu@fas.harvard.edu

Abstract

The study of arrangements of non-overlapping spheres in space, known as sphere packings, has given rise to numerous questions such as finding the densest sphere packing and the kissing number problem. Aside from these theoretical considerations, sphere packings is also closely related to the theory of codes. This paper introduces the basic problem of finding efficient error-correcting codes and discusses its geometric interpretation as a sphere packing problem. The paper focuses on certain families of codes with special properties and explores these properties in connection with two famous codes.

1.1 Introduction

Consider the n-dimensional real vector space $\mathbb{R}^n$ equipped with an inner product $\langle \cdot, \cdot \rangle$ and the metric $|v| = \sqrt{\langle v, v \rangle}$. We ask: how can we embed non-overlapping unit spheres in $\mathbb{R}^n$ in order to maximize the density of the arrangement, that is, the proportion of space filled by the spheres? In examining such arrangements, called **sphere packings**, one can consider a special type of arrangement, in which the centers of the spheres form a special structure called a lattice.

A **lattice** is the abelian group $\mathbb{Z}^n$ equipped with a homomorphism $\varphi : \mathbb{Z}^n \xrightarrow{f} \mathbb{R}^n$ such that the images of the standard basis vectors $\{e_i\} \in \mathbb{Z}^n$ form a basis of $\mathbb{R}^n$. Such a lattice is said to be **of rank n**. As an example, consider the rank-2 hexagonal lattice, pictured in Figure 1.1. We define $f(e_1) = (1, 0)$ and $f(e_2) = \frac{1}{2}(1, \sqrt{3})$:

If we arrange spheres with centers located at each of the hexagonal lattice points, we obtain a **sphere packing** in $\mathbb{R}^2$. Carl Friedrich Gauss famously proved that this is the densest lattice packing in $\mathbb{R}^2$, having a density of $\frac{\pi}{2\sqrt{3}} \approx 0.9069$. The hexagonal packing was later proven to be the densest sphere packing in $\mathbb{R}^2$.

While sphere packings are interesting mathematical objects of study by themselves, important applications of sphere packings arise in the design of signals for use in data transmission. In a typical system, there is an **information source**, such as a human speaker, which produces messages to be communicated to a destination. These messages are then converted to digital form via a **source encoder** and transmitted across a **noisy channel**, such as a copper wire or radio waves. The

[†]Francois Greer '11 is a mathematics concentrator and physics minor. His interests lie mainly in algebra and related topics, and he currently serves as Features Editor of The HCMR.

[‡]Xiaoqi Zhu '11 is a mathematics concentrator. Outside of his mathematical interests, he also studies economics and plans to minor in government.

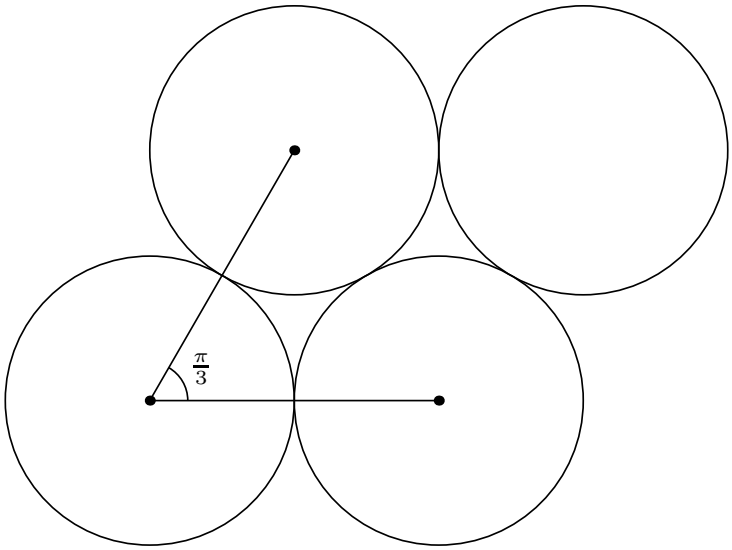


Figure 1.1: The hexagonal lattice maximally packs spheres in $\mathbb{R}^2$.

noise in the channel inevitably scrambles the signal to some extent. In order to recover the original message from the decoder, we wish to design a set of special signals in which errors can be detected and corrected. An **error-correcting code** is one such set of special signals, whose members are designed to be easily distinguishable from each other even in the presence of noise. Thus, error-correcting codes help prevent miscommunication of a message by correcting the scrambling that occurs during transmission. When an error-correcting code is implemented, the **channel encoder** takes the output from the source encoder, replaces it with a corresponding code signal, and transmits the code signal over the noisy channel. The **channel decoder** subsequently attempts to recover the original message by taking the received signal and estimating the sent code signal. Using error-correcting codes, we can communicate messages from a source to a destination with greater reliability. As we will see, the existence of efficient error-correcting codes is closely related to dense sphere packings.

In this article, we explore various error-correcting codes and their connections to sphere packings. We begin by formally defining error-correcting codes and providing intuition for the coding problem. In Section 1.3, we provide several examples of codes. Sections 1.4 and 1.5 explore certain families of codes with special properties, which are subsequently used to define and explain two particular error-correcting codes: the Hamming code and the binary Golay code. Section 1.8 explains several similarities between the Hamming and Golay codes via the theory of quadratic residue codes. In Section 1.9, we establish the connection between error-correcting codes and Euclidean sphere packings. In particular, we show that the extended Hamming and extended binary Golay codes can be used to construct the densest lattice packings in dimensions 8 and 24, respectively. We conclude by explaining the concept of perfect codes and providing a full characterization of such codes.

1.2 Error-correcting Codes

A **code** C of length n is a set of vectors called **codewords** in $\mathbb{F}_q^n$ where $\mathbb{F}_q$ is the Galois field of order $q = p^r$, for p prime. As a subset of a finite vector space, a code is a finite set. A **linear code** C is a linear subspace of $\mathbb{F}_q^n$, and is thus closed under vector addition and coordinatewise multiplication by elements of $\mathbb{F}_q$. For this paper, we will restrict our attention to **binary linear codes**, *i.e.* subspaces of $\mathbb{F}_2^n = \{0,1\}^n$. **Channel noise** is a probability $p < 1/2$ that, when a 0 or 1 is sent in a codeword, a different symbol is received by the decoder. Given a code C, we can define a natural metric called the **Hamming distance** $d(u,v)$ between two vectors $u = (u_1, \ldots, u_n)$ and

$v = (v_1, \ldots, v_n)$ in C:

$$d(u, v) = |\{i : u_i \neq v_i\}|.$$

The Hamming distance is the number of coordinates in which u and v differ. This clearly satisfies the axioms of a metric. The **Hamming weight** of a vector $u = (u_1, \ldots, u_n)$ in C is its distance from the zero vector, or

$$wt(u) = |\{i : u_i \neq 0\}|.$$

Clearly, $d(u, v) = wt(u - v)$. The **minimal distance** d of a code C is defined as

$$d = \min\{d(u, v) : u, v \in C, u \neq v\}.$$

For linear codes, $d = wt(u)$, where u is the non-zero codeword of minimal weight. A linear code of length n, dimension k and minimal distance d is called an $[n, k, d]$ code.

Given d, the corresponding **Hamming radius** is defined as

$$\rho = \frac{1}{2}(d - 1).$$

We can construct disjoint closed balls of radius ρ around every codeword in a code of minimal distance d. In this sense, we can regard a code as a packing of spheres in $\mathbb{F}_2^n$ with radius ρ and centers at each point $c \in C$. In Section 1.9, we will show how a code can be extended to yield a sphere packing in $\mathbb{R}^n$.

This geometric interpretation serves as a useful method for understanding the process of signal transmission and error correction. When a message of length k is encoded using an $[n, k, d]$ code with Hamming radius ρ, it is replaced by a codeword in $\mathbb{F}_2^n$, $n \geq k$. During the process of transmission, the signal is scrambled to some extent, but as long as the scrambled signal falls within the Hamming radius of some codeword, the decoder is able to "correct" the signal by sending it to the center of the Hamming sphere containing it, *i.e.* the corresponding codeword.[1] The code thus corrects ρ errors, since it corrects any received signal for which there exists a codeword different in no more than ρ digits from the signal. Geometrically, correctable signals fall within Hamming distance ρ of some codeword. One might be tempted to conclude that a good error-correcting code has large minimal distance so as to maximize ρ. However, such codes are not necessarily effective.

To see this, consider the following code consisting of just two codewords:

$$C = \{(\overbrace{0, \ldots, 0}^{n \text{ times}}), (\overbrace{1, \ldots, 1}^{n \text{ times}})\}.$$

This is called the **repetition code** of length n. The repetition code takes every bit of information and passes it through the channel as the same bit repeated n times. Because the code has minimal distance n, it has the maximal Hamming radius of a length n code. From the sole standpoint of error correction, one might say that the repetition code is effective. In particular, if n is odd, then every signal falls within the Hamming radius of a codeword. Such a code is thus able to recover a codeword from every possible signal scrambling. However, the repetition code is slow to implement because it can only encode one bit at a time. Moreover, for the code to work with reasonable accuracy, n must be large.

Thus, an effective error-correcting code must have a large number of codewords to allow longer messages to be encoded. Additionally, the length of the code must be small, to ensure efficient processing. Thus, a good code is one with n small (to reduce delays), k large (to increase message versatility), and d large (for better error-correcting). These are naturally incompatible goals, hence the construction of good error-correcting codes is an interesting and difficult problem.

[1] This does not ensure that the signal is corrected properly; the resulting codeword can be different from the original codeword.

1.3 Examples of Codes

We can better understand the concepts presented in the previous section by looking at some additional examples of codes:

Example 1. An $[n, 0, n]$ code is called a **zero code** of length n. Such a code consists of simply the zero vector $(0, \ldots, 0) \in \mathbb{F}_2^n$. It has dimension 0 and minimal distance n.

Example 2. An $[n, n, 1]$ code is called a **universe code** of length n. Such a code consists of all vectors $(c_1, \ldots, c_n) \in \mathbb{F}_2^n$. It has dimension n (since it is equal to $\mathbb{F}_2^n$) and minimal distance 1.

Example 3. An $[n, 1, n]$ code is called a **repetition code** of length n. Such a code consists of the vectors $(0, \ldots, 0)$ and $(1, \ldots, 1)$. It has dimension 1 (since it is isomorphic to $\mathbb{F}_2$) and minimal distance n.

Example 4. An $[n, n-1, 2]$ code is called a **zero-sum code** of length n. Such a code consists of all vectors $(c_1, \ldots, c_n) \in \mathbb{F}_2^n$ such that $\sum c_i = 0$. It has dimension $n-1$, which we can see explicitly by considering the basis

$$\{(1, 1, 0, \ldots, 0, 0), (0, 1, 1, \ldots, 0, 0), \cdots, (0, 0, 0, \ldots, 1, 1)\}$$

for the code. It has minimal distance 2 since a vector $(c_1, \ldots, c_n)$ with only one non-zero digit has $\sum c_i = 1 \neq 0$.

1.4 Dual Codes

Given a linear code C, we define the **dual code** C^* of C to be the orthogonal complement of C in $\mathbb{F}_2^n$ with respect to the dot product. More formally,

$$C^* = \{v \in \mathbb{F}_2^n : v \cdot c = 0 \; \forall c \in C\}.$$

It can be easily shown that $\dim C + \dim C^* = n$, and that $(C^*)^* = C$. A code is said to be **self-dual** if $C^* = C$.

Revisiting our previous examples, we see that the universe code of length n (Example 2) is the dual code of the zero code of length n (Example 1), because the dot product of the zero vector with any vector in $\mathbb{F}_2^n$ vanishes. It is also easy to see that the zero-sum code of length n (Example 4) is the dual code of the repetition code of length n (Example 3), because the dot product of $(0, \ldots, 0)$ with any vector vanishes, and

$$(1, \ldots, 1) \cdot (c_1, \ldots, c_n) = 0 \Leftrightarrow \sum c_i = 0.$$

1.5 Cyclic Codes

A code C is **cyclic** if $(c_0, \ldots, c_{n-2}, c_{n-1}) \in C$ implies that $(c_{n-1}, c_0, \ldots, c_{n-2}) \in C$. It is convenient to represent a cyclic code of length n as a polynomial quotient ring. A codeword $c = (c_0, \ldots, c_{n-1}) \in C$ can be written as a polynomial:

$$c(x) = c_0 + c_1 x + \ldots + c_{n-2} x^{n-2} + c_{n-1} x^{n-1}.$$

Consider the ring $R = \mathbb{F}_2[x]/(x^n - 1)$, where $(x^n - 1)$ is the ideal generated by the polynomial $x^n - 1$. Then the cycling action corresponds to multiplication by x in R:

$$\begin{aligned} xc(x) &= c_0 x + c_1 x^2 + \ldots + c_{n-2} x^{n-1} + c_{n-1} x^n \\ &\sim c_0 x + c_1 x^2 + \ldots + c_{n-2} x^{n-1} + c_{n-1} x^n - c_{n-1}(x^n - 1) \\ &= c_{n-1} + c_0 x + c_1 x^2 + \ldots + c_{n-2} x^{n-1}. \end{aligned}$$

A cyclic code corresponds to an ideal of R because it is closed under (polynomial) addition and under cycling, or multiplication by x. Since $\mathbb{F}_2[x]$ is a principal ideal domain, there exists a **generating polynomial** $g(x)$ for C. From a computational standpoint, an interesting problem is to

represent C as a polynomial field, *i.e.* the quotient of $\mathbb{F}_2[x]$ by a maximal ideal (an ideal generated by an irreducible polynomial). While an irreducible polynomial of degree n always exists, working with such a polynomial is often rather cumbersome. Good candidates for computationally simple irreducibles are trinomials of the form $x^n + ax + b$. It remains an open problem in mathematics to determine whether such irreducible trinomials exist over $\mathbb{F}_2$ for all n.

Returning to the above examples, the universe code has generator $g(x) = 1$. The repetition code has generator $g(x) = 1 + x + \ldots + x^{n-1}$. The zero-sum code has generator $g(x) = 1 + x$, which incidentally corresponds to our choice of basis. Note that these generators are not unique, but can nonetheless be useful for defining more complex codes.

1.6 Hamming Codes

The **Hamming code** $\mathcal{H}_7$ is a binary cyclic code of length 7 with generating polynomial $g(x) = 1 + x + x^3$. By definition, it is spanned by the set

$$\begin{aligned} G = \{&(1,1,0,1,0,0,0),\\ &(0,1,1,0,1,0,0),\\ &(0,0,1,1,0,1,0),\\ &(0,0,0,1,1,0,1),\\ &(1,0,0,0,1,1,0),\\ &(0,1,0,0,0,1,1),\\ &(1,0,1,0,0,0,1)\}. \end{aligned}$$

The first four vectors in G are linearly independent, and we can obtain the latter three vectors by linear combination of the first four vectors. Thus, the Hamming code is a 4-dimensional linear subspace of $\mathbb{F}_2^7$ and has $2^4 = 16$ codewords. It can be checked that the Hamming code has minimal distance 3, *i.e.* any two codewords differ in at least three coordinates. Thus, the Hamming code is a $[7, 4, 3]$ code.

Given a code C, we can "extend" the code by appending a digit to the end of every codeword such that the sum of the digits of each codeword is zero. This digit is called a zero-sum check digit. Formally, given a code C, the extended code is defined as

$$C' = \left\{ (c_1, \ldots, c_n, c_{n+1}) : \sum_{i=1}^{n+1} c_i = 0, (c_1, \ldots, c_n) \in C \right\}.$$

The **extended Hamming code** $\mathcal{H}_8$ is a binary cyclic code of length 8 and dimension 4 generated by the basis

$$\begin{aligned} G' = \{&(1,1,0,1,0,0,0,1),\\ &(0,1,1,0,1,0,0,1),\\ &(0,0,1,1,0,1,0,1),\\ &(0,0,0,1,1,0,1,1)\}. \end{aligned}$$

Notably, the coordinates of every codeword in $\mathcal{H}_8$ consists of precisely four 0's and four 1's, with the exception of the codewords $(0,0,0,0,0,0,0,0)$ and $(1,1,1,1,1,1,1,1)$. It follows immediately that the extended Hamming code is **doubly even**, *i.e.* the sum of all coordinates in every codeword is divisible by four. Moreover, the extended Hamming code has minimal distance 4, and is therefore an $[8, 4, 4]$ code. It can be checked that the extended Hamming code is self-dual.

Since the extended Hamming code has minimal distance 4, it has a Hamming radius of $\rho = \frac{3}{2}$. The extended Hamming code can therefore correct all single-bit errors. Richard Hamming published the original $[7, 4, 3]$ Hamming code in 1950. At the time, existing error-correcting codes were either inefficient (such as the repetition code) or unrobust in their error-correction (such as Bell's 'two-out-of-five' code). Hamming searched for a code that would have both a large Hamming distance and a fast information rate. The result was the $[7, 4, 3]$ Hamming code.

1.7 Golay Codes

The **binary Golay code** $\mathcal{G}_{23}$ is a binary cyclic code of length 23 with generating polynomial $g(x) = 1 + x^2 + x^4 + x^5 + x^6 + x^{10} + x^{11}$. It can be checked that the Golay code is a $[23, 12, 7]$ code.

The **extended binary Golay code** $\mathcal{G}_{24}$, constructed by appending a zero-sum check digit to the end of each codeword in $\mathcal{G}_{23}$, is a 12-dimensional linear subspace of $\mathbb{F}_2^{24}$ and has $2^{12} = 4096$ codewords. All codewords in the extended binary Golay code have Hamming weight 0, 8, 12, 16 or 24. This implies that the extended binary Golay code is doubly even with minimal distance 8 (and is therefore a $[24, 12, 8]$ code). It can be checked that the extended binary Golay code is self-dual.

Since the extended binary Golay code has minimal distance 8, it has a Hamming radius of $\rho = \frac{7}{2}$. The extended binary Golay code can therefore correct all triple-bit errors. The original $[23, 12, 7]$ Golay code, discovered by Marcel Golay in 1949, is the only error-correcting code known capable of correcting any combination of three or fewer random errors in a sequence of 23 elements.

As $\mathcal{G}_{24}$ is a high-dimensional code, it can be practically applied with reasonable efficiency. During the 1979 and 1980 NASA Deep Space Missions, the Voyager 1 and Voyager 2 spacecrafts needed to transmit hundreds of high-definition color images of Jupiter and Saturn. Transmission of these images involved communicating a large amount of data via a constrained telecommunications bandwidth. In order to transmit the data efficiently, NASA implemented the Golay code. Although there were codes that could correct greater errors (such as the $[32, 6, 16]$ Hadamard code, which corrects up to 7-bit errors), the Golay code was preferred because it yields a much higher data rate.

1.8 Quadratic Residue Codes

Let p be an odd prime. We define the set of **quadratic residues modulo p** as

$$Q := \{n^2 \bmod p : n \in \mathbb{F}_n, n \neq 0\}.$$

It is easy to see that Q forms a multiplicative group, since $a^2b^2 = (ab)^2$ and $(a^2)^{-1} = (a^{-1})^2$. We can use this idea to construct an important class of cyclic codes called **quadratic residue codes**, which reveal a connection between the Hamming and Golay codes. In order to do this, pick p such that 2 is a quadratic residue modulo p. Let ζ be a primitive p-th root of unity in some finite extension field $K \supset \mathbb{F}_2$.

Lemma 5. *The polynomial* $g(x) = \prod_{j \in Q} (x - \zeta^j)$ *has coefficients in* $\mathbb{F}_2$.

Proof. Let us write

$$g(x) = \sum_{k=0}^{n} a_k x^k.$$

By construction, it is clear that each a_k is a symmetric polynomial in the elements $\zeta^j \in K$. We claim that $a_k^2 = a_k$. In a field of characteristic 2, $(\alpha + \beta)^2 = \alpha^2 + \beta^2$ (this is a simple computational exercise). Hence, the operation of squaring permutes elements in the symmetric polynomial via $(\zeta^j)^2 = \zeta^{2j}$ (2 is a quadratic residue so $2Q = Q$). Thus we have $a_k^2 = a_k \Rightarrow a_k = 0$ or $1 \Rightarrow g(x) \in \mathbb{F}_2[x]$. □

The cyclic code of length p generated by $g(x)$ is known as the **quadratic residue code $\mathcal{Q}_p$**.

A well-known theorem of number theory states that the primes p for which 2 is a quadratic residue must be of the form $8k \pm 1$. Thus we can construct quadratic residue codes for $p = 7, 17, 23, \ldots$ All such codes have dimension $(p+1)/2$ and minimal distance $d \geq \sqrt{p}$. Furthermore, if $p = 8k - 1$ then the code obtained by appending a zero-sum check digit to the codewords in $\mathcal{Q}_p$ is self-dual. The Hamming and Golay codes are two low examples of quadratic residue codes, which is why they share several remarkable properties.

1.9 Euclidean Sphere Packing Construction

As previously demonstrated, a binary code C of length n corresponds to a sphere packing in $\mathbb{F}_2^n$. We can additionally construct from every code a sphere packing in $\mathbb{R}^n$ with centers at each point $x = (x_1, \ldots, x_n)$ such that $x \equiv c \bmod 2$ for some $c \in C$. A lattice packing is obtained if and only if C is linear.

We can apply this construction to the extended Hamming code and the extended Golay code to generate some remarkable lattices. From $\mathcal{H}_8$, one can construct a lattice $H_8 \simeq E_8$, where

$$E_8 = \left\{ (x_i) \in \mathbb{Z}^8 \cup \left(\mathbb{Z} + \frac{1}{2}\right)^8 : \sum_{i=1}^{8} x_i \equiv 0 \bmod 2 \right\}.$$

That is, E_8 is the set of points in $\mathbb{R}^8$ such that either all coordinates are integers or are half-integers, and the sum of all coordinates is an even integer. Notably, E_8 is the densest lattice packing in $\mathbb{R}^8$. Hence, the sphere packing construction applied to $\mathcal{H}_8$ yields a lattice packing isomorphic to the densest lattice packing in $\mathbb{R}^8$.

Even more remarkably, the construction can be applied to the extended binary Golay code to generate a lattice isomorphic to the Leech lattice Λ_{24}, the densest lattice packing in $\mathbb{R}^{24}$. For more information regarding these constructions as well as E_8 and Λ_{24}, we refer the reader to [CS, ch. 5].

1.10 Perfect Codes

When a codeword is transmitted, it may be scrambled in such a way that the resulting signal no longer falls within the Hamming radius of any codeword. This problem is resolved by perfect codes. A code C of length n is **perfect** if every element of $\mathbb{F}_2^n$ is contained in a Hamming sphere about one of the codewords. A perfect code with Hamming radius ρ thus has the property that every received signal lies within a Hamming distance of ρ from exactly one codeword. In essence, every received signal has a well-defined message. In the language of sphere packings, perfect codes are sphere packings in $\mathbb{F}_2^n$ with maximal density 1. Perfect codes have been classified into four categories:

1. Trivial codes (universe codes, zero code, repetition codes of odd length)
2. Hamming codes ($\mathcal{H}_7$ and $\mathcal{H}_8$)
3. Nonlinear binary codes (not fully enumerated)
4. Golay codes ($\mathcal{G}_{23}$ and $\mathcal{G}_{11}$)

In this paper, we examined all perfect linear codes with the exception of the ternary Golay code $\mathcal{G}_{11}$, a 6-dimensional linear subspace of $\mathbb{F}_3^{11}$ with minimal distance 5. Currently, the only known perfect non-binary linear code is $\mathcal{G}_{11}$.

We now derive a proposition useful for identifying perfect linear codes. The **Hamming bound** can be stated as follows:

Theorem 6. *For an $[n, k, d]$ code C with Hamming radius $\rho = \frac{1}{2}(d-1)$,*

$$2^n \geq 2^k \sum_{i=0}^{\lfloor \rho \rfloor} \binom{n}{i}.$$

Proof. C has 2^k codewords. For each codeword $c \in C$, there are $\binom{n}{i}$ elements in $\mathbb{F}_2^n$ that differ from c in exactly i positions. Thus, $2^k \sum_{i=0}^{\lfloor \rho \rfloor} \binom{n}{i}$ is simply the total number of vectors in $\mathbb{F}_2^n$ that lie within a Hamming distance of ρ of some codeword. Since this number cannot exceed the total number of vectors in $\mathbb{F}_2^n$, we obtain the desired inequality. □

Clearly, an $[n, k, d]$ code C is perfect if and only if the Hamming bound is an equality. Therefore, the Hamming bound provides a method for verifying perfect codes. For $\mathcal{H}_7$, we have

$$2^7 = 2^4 \left[\binom{7}{0} + \binom{7}{1} \right].$$

For $\mathcal{G}_{23}$, we have

$$2^{23} = 2^{12} \sum_{i=0}^{3} \binom{23}{i}.$$

Proofs that $\mathcal{H}_8$, $\mathcal{G}_{11}$ and the so-called "trivial codes" are perfect follow in the same fashion.

Unlike $\mathcal{G}_{23}$, the extended binary Golay code is not perfect. However, it can be verified that every vector in $\mathbb{F}_2^{24}$ lies within a Hamming distance $\rho + 1$ of some codeword. Such codes are known as quasi-perfect codes.

Acknowledgements

Much of the material in this article was provided in or motivated by Professor Benedict Gross's freshman seminar, "Euclidean Lattices and Sphere Packings," in Spring 2008.

References

[CS] Conway, J.H. and Sloane, N.J.A.: *Sphere Packings, Lattices and Groups*. New York: Springer 1998.

[Ha] Hamming, R.W.: Error Detecting and Error Correcting Codes, *The Bell System Technical Journal* **26** #2 (1950), 147–160.

[vL] van Lint, J.H.: *Introduction to Coding Theory*, Graduate Texts in Mathematics. Berlin: Springer-Verlag 1999.

STUDENT ARTICLE

2 Kummer, Regular Primes, and Fermat's Last Theorem

Ila Varma[†]
California Institute of Technology '09
Pasadena, CA 91126
ila@caltech.edu

Abstract

This paper rephrases Kummer's proof of many cases of Fermat's Last Theorem in contemporary notation that was in fact derived from his work. Additionally, this paper develops a reformulation of the proof using class field theory from a modern perspective in a manner similar to the tactics used for the complete proof, and describes how Kummer's proof strategy can generalize to solve the theorem for a broader set of primes.

2.1 Introduction

Ernst Kummer was a 19th century mathematician who came across Fermat's Last Theorem in attempts to generalize the law of quadratic reciprocity and study higher reciprocity laws. While he described those as "the principal subject and the pinnacle of contemporary number theory," he considered Fermat's Last Theorem a "curiosity of number theory rather than a major item" ([Ed]). *A priori*, this was not an unreasonable opinion of a problem that could be understood by a 12-year-old. We state this mere curiosity below.

Theorem 1. *For any integer $n > 2$, the equation $x^n + y^n = z^n$ has no non-trivial solutions in the integers, i.e. if $x, y, z \in \mathbb{Z}$ satisfy this equation, then $xyz = 0$.*

Despite his disinterest, Kummer made the first substantial step in proving a part of Fermat's Last Theorem for many cases. This came only a few weeks after Gabriel Lamé incorrectly announced that he had found a complete proof [Ed]. Lamé did make the breakthrough in attempting to decompose $x^n + y^n$ into linear factors by introducing the complex numbers satisfying $\zeta^n = 1$, known today as roots of unity. This allowed for the algebraic identity

$$x^n + y^n = (x + y)(x + \zeta y)(x + \zeta^2 y) \cdots (x + \zeta^{n-1} y).$$

Thinking this was the only new step needed to find the complete solution, Lamé presented a proof in March 1847 using this fact while assuming incorrectly that this was a unique decomposition into prime ideals [Ed]. A few years earlier, Kummer had already discovered that such unique factorization properties did not necessarily hold in the fields $\mathbb{Q}(\zeta_p)$ generated by these roots of unity. He introduced the origins of the notion of an ideal in an attempt to salvage the absence of unique factorization, as well as the class number and an analytic formula describing it [Ri]. A few weeks after Lamé presented his incorrect proof, Kummer wrote a correct proof for a certain set of prime. These primes had a property allowing for unique factorization to work in the step of Lamé's proof that went wrong. He called these regular primes, and in his later work, continued his examination of both regular and non-regular primes to find straightforwards characterizations and deeper properties. In his proof and this further examination, Kummer touched on ideas that would be developed into present-day ideal theory, Kummer theory, p-adic analysis, class field theory,

[†]Ila Varma is currently a senior studying mathematics at California Institute of Technology. Her research interests fall into the areas of number theory and algebraic geometry.

etc. The core ideas behind modern problems such as the Birth-Swinnerton-Dyer conjecture for the complex multiplication case and the theorems of Clausen and von Staudt are influenced by Kummer's work, not to mention the ideas that led to the eventual complete solution of Fermat's Last Theorem ([vdP]).

In this article we will focus on Kummer's ideas regarding and influence on the solution of Fermat's Last Theorem, and thus we will stay in the realm of proving the theorem for regular primes. Many of the preceding lemmas are given in detail as a demonstration of the machinery needed, but additionally, a modern perspective is described, along with the generalization of Kummer's idea to a larger set of primes. Section 2.2 gives a background on cyclotomic fields and describes some properties needed for the proof based on Kummer's original work described in Sections 2.3 and 2.4. Section 2.5 reformulates the proof using an approach matching the strategies used for the complete solution. Finally, Section 2.6 is devoted to proving Fermat's Last Theorem for the most general characterization of primes on which Kummer's basic argument holds.

2.2 Background

We must first describe general notation and some basic facts on cyclotomic fields and algebraic number theory. Then, we can go on to understand the core idea from the proof, and in particular where the regularity of primes fits in and therefore restricts the cases of Fermat's Last Theorem.

Roots of Unity and Cyclotomic Fields

For any odd prime p, we denote by ζ_p a fixed **primitive p-th root of unity**, *i.e.* a $\zeta_p \in \mathbb{C}$ such that $\zeta_p^k \neq 1$ for any $k = 1, \ldots, p-1$ while $\zeta_p^p = 1$. It, along with all of its powers, is a root of the polynomial $x^p - 1$, hence it satisfies the equation $x^p = 1$, the motivation for its name. To find its minimal polynomial, we note that the only rational p-th root of unity is $\zeta_p^p = 1$, hence we can factor $x^p - 1 = (x-1)\Phi_p(x)$ where

$$\Phi_p(x) = \frac{x^p - 1}{x - 1} = x^{p-1} + x^{p-2} + \cdots + x + 1.$$

This is called the **p-th cyclotomic polynomial** as it is the minimal polynomial for ζ_p. Note that the other p-th roots of unity are powers of ζ_p, and are all roots of $\Phi_p(x)$ (except for $\zeta_p^p = 1$). From an analytic perspective, we can think of $\zeta_p = e^{2\pi i/p}$, and subsequently, $\zeta_p^k = e^{2\pi ik/p}$. Here, we can see that all powers of ζ_p lie on unit circle ($f(z) = e^{2\pi iz}$) in the complex plane, and furthermore, the shape described with ζ_p^k as vertices is a regular p-gon. Furthermore in $\mathbb{C}$, we can factor $\Phi_p(x) = \prod_{k=1}^{p-1}(x - \zeta_p^k) = x^{p-1} + x^{p-2} + \cdots + x + 1$. From here, we know that the product of the non-trivial p-th roots of unity (*i.e.* not including 1) has magnitude 1 (the constant coefficient) and the sum of the non-trivial p-th roots of unity also has magnitude 1 (the x^{p-2} coefficient). More explicitly, we have the relation

$$\zeta_p + \zeta_p^2 + \zeta_p^3 + \cdots + \zeta_p^{p-1} = -1.$$

Hence, we have that the sum of all of the p-th roots of unity is 0, *i.e.* any p-th root of unity can be expressed as a linear sum of its other powers. It is in fact true that that any set of $p-1$ roots of unity are linearly independent while the whole set is not.

We can also talk about the field generated by p-th roots of unity over $\mathbb{Q}$ known as the **p-th cyclotomic field**. Note that this field, denoted $K = \mathbb{Q}(\zeta_p)$, is automatically the splitting field for $\Phi_p(x)$ over $\mathbb{Q}$ as we have seen before that the rest of the roots are just subsequent powers of ζ_p. This extension has degree $p-1$, coinciding with the degree of $\Phi_p(x)$. Furthermore, the group of automorphisms well-defined on K that fix $\mathbb{Q}$ is cyclic. More explicitly, the Galois group, $\mathrm{Gal}(K/\mathbb{Q}) \cong (\mathbb{Z}/p\mathbb{Z})^\times$ where the automorphism $\left[\sigma_k : \zeta_p \mapsto \zeta_p^k\right] \mapsto k$.

A Bit of Algebraic Number Theory

Stepping back from the specifics of cyclotomic fields, we can realize many useful properties of **number fields**, *i.e.* algebraic extensions over $\mathbb{Q}$ from algebraic number theory. First note that any

monic polynomial with coefficients in $\mathbb{Z}$ that is known to have roots in $\mathbb{Q}$ in fact has roots in $\mathbb{Z}$ (this is the Rational Root Theorem from high-school algebra). This interesting property can be used to describe the structure of $\mathbb{Z}$ within $\mathbb{Q}$, and we generalize this to any number field K. The elements of K which are roots of monic polynomials with coefficients in $\mathbb{Z}$ are known as the **algebraic integers** of K and furthermore produce a **ring of integers**, generally denoted $\mathcal{O}_K$. As an example, the ring of integers of any p-th cyclotomic field $\mathbb{Q}(\zeta_p)$ is $\mathbb{Z}[\zeta_p]$.

Like any other ring, $\mathcal{O}_K$ has ideals, and one property is that the ring of integers for any field K is a **Dedekind domain**, a type of integral domain with the added property that any ideal decomposes uniquely into a product of prime ideals. It is not necessarily true, however, that the elements of a Dedekind domain decompose uniquely into a prime or irreducible elements. It is not hard to find an example displaying this unfortunate fact: if $K = \mathbb{Q}(\sqrt{-5})$, then $\mathcal{O}_K = \mathbb{Z}[\sqrt{-5}]$, and we can consider $6 = 2{\cdot}3 = (1+\sqrt{-5})\cdot(1-\sqrt{-5})$. Nevertheless, we can see that if all the ideals of a given $\mathcal{O}_K$ are principal, then the unique decomposition of prime ideals would give way to unique prime factorization of elements, as the factorization of any element $\alpha \in \mathcal{O}_K$ would be characterized by the decomposition of the ideal (α) into prime ideals generated by single irreducible elements. This motivates the construction of the **ideal class group** of K which is, loosely speaking, the quotient group of all the ideals in $\mathcal{O}_K$ modulo the principal ideals of $\mathcal{O}_K$. We are very lucky to find that this group is always finite, and in fact, when the order is 1, we are in the previously-described case, in which all ideals of $\mathcal{O}_K$ are principal. The **class number** of K, denoted h_K, is the order of this ideal class group. Hence, if $h_K = 1$, $\mathcal{O}_K$ has unique prime factorization of elements. If $h_K > 1$, then $\mathcal{O}_K$ does not have unique prime factorization, but we can be more specific than that. The class number does describe the extent to which unique factorization holds; for example, there are properties about length of decompositions in fields of class number 2 that do not hold for fields with higher class number.

Regular Primes

The property of whether a prime p is **regular** can be characterized based on the class number of $K = \mathbb{Q}(\zeta_p)$. Explicitly, the class number h_K is the order of the ideal class group, but as described above, we think of the class number as a scalar quantity describing how "close" elements of $\mathcal{O}_K$ are to having unique factorization.

Definition 2. An odd prime p is **regular** if the class group of $K = \mathbb{Q}(\zeta_p)$ has no p-torsion, *i.e.* if the class number h_K is prime to p.

It is astonishing to think that such a fact should be related to the ease of proving Fermat's Last Theorem, but it is in fact the case. Lamé's first step of decomposing a nontrivial counterexample $x^p + y^p = z^p$ in the field $\mathbb{Q}(\zeta_p)$ only goes so far when we don't have unique prime factorization of the elements. It is easy to work with z^p when considering it as an ideal of $\mathcal{O}_K$, but at some point, we must be able to look at specific elements of the ring of integers rather than the ideals they generate. *A priori* as ideals, we get

$$(z)^p = (x+y)(x+\zeta_p y)(x+\zeta_p^2 y)\cdots(x+\zeta_p^{p-1} y).$$

On one side, we have a p-th power of a principal ideal, and on the other, we have a decomposition into p ideals that are not only distinct, but can be shown to be relatively prime. The property of unique decomposition into prime ideals tells us that every ideal $(x+\zeta_p^k y)$ must independently be a p-th power of an ideal A_k. Thinking about the structure of the ideal class group, we can consider what kinds of ideals of $\mathcal{O}_K$ have a p-th power which is principal. As elements of this quotient group, the ideals of $\mathcal{O}_K$ modulo the principal ideals of $\mathcal{O}_K$, it is clear that $(x+\zeta_p^k y)$ is identified with the trivial element, but it is not necessarily true that A_k is. Nevertheless, if we have the added assumption that the prime p is regular, then we know that the class number is prime to p, hence no element in the ideal class group can have order p without being trivial. This directly implies that A_k is principal, and we can in fact think about the element α_k generating this ideal rather than the ideal $(\alpha_k) = A_k$ itself. From this point onward, Kummer's proof consists of algebraic manipulations of units and algebraic integers in K leading to a contradiction that cannot be done simply by working with ideals. It is easy to see that the regularity of p is the broadest way to guarantee that the p-th

"roots" of the ideals generated by $x+\zeta_p^k y$ are in fact principal, bringing the entire machinery down to elements of $\mathcal{O}_K$.

Prime Decomposition

It is interesting to note that prime ideals of a base field may not stay prime in an extension. For example, we can show that in $K = \mathbb{Q}(\zeta_p)$, the ideal generated by p decomposes as

$$(p) = (1-\zeta_p)^{p-1}.$$

where $(1-\zeta_p)$ turns out to be a prime ideal of $\mathcal{O}_K$. In algebraic number theory, we are quite interested in *how* a prime ideal such as (p) in a base field $\mathbb{Q}$ occurs in a larger field such as K. It is "easiest" when a prime stays **inert**, *i.e.* stays prime in the larger field extension. However, in many cases, such as the above, a prime ideal of the base field will decompose further in the extension, and it is particular interesting to note when such a decomposition includes repeated factors, *i.e.* when the prime **ramifies**. The case where the prime of the base field can be written as a power of a single prime ideal in the extension is known as **total ramification.** Additionally, for total ramification, we require that the power of the single ideal coincides with the degree of the extension. As an example, we will prove that p totally ramifies in K as a power of $(1-\zeta_p)$. To prove the above fact about (p), we first introduce the notion of cyclotomic units.

Definition 3. The **cyclotomic units** of $\mathcal{O}_K = \mathbb{Z}[\zeta_p]$ are elements of the form

$$\frac{\zeta_p^r - 1}{\zeta_p^s - 1} \quad \text{where } p \nmid rs.$$

It is easy to see that these are units as they have obvious inverses, $\frac{\zeta_p^s-1}{\zeta_p^r-1}$, hence the cyclotomic units are in fact a subgroup of $\mathcal{O}_K^\times$. Furthermore, we see that since $p \nmid rs$, we can find t such that $r = st \bmod p$ hence allowing us to express

$$\frac{\zeta_p^r - 1}{\zeta_p^s - 1} = \frac{\zeta_p^{st} - 1}{\zeta_p^s - 1} = 1 + \zeta_p^s + \cdots + \zeta_p^{s(t-1)} \in \mathcal{O}_K.$$

Lemma 4. *The principal ideal generated by p in $\mathcal{O}_K$ decomposes as $(1-\zeta_p)^{p-1}$, and hence the principal ideal $(1-\zeta_p)$ is prime in $\mathcal{O}_K$.*

Proof. Since the minimal polynomial of ζ_p is $\Phi_p(x) = \frac{x^p-1}{x-1}$, as a polynomial in $K[x]$, it can be decomposed as

$$\Phi_p(x) = \prod_{i=1}^{p-1}(x - \zeta_p^i).$$

Note that if we plug in $x = 1$ to $\Phi_p(x)$ we get from the polynomial in $\mathbb{Q}[x]$ and the polynomial in $K[x]$ that

$$p = \prod_{i=1}^{p-1}(1 - \zeta_p^i).$$

Note that $1-\zeta_p$ is a unit away from $1-\zeta_p^i$, *i.e.* $1-\zeta_p^i = u(1-\zeta_p)$ where u is the cyclotomic unit $\frac{\zeta_p^i-1}{\zeta_p-1}$. Thus we have an equality of ideals $(1-\zeta_p) = (1-\zeta_p^i)$. This, combined with the decomposition of p gives us $(p) = (1-\zeta_p)^{p-1}$. Furthermore, since $[K:\mathbb{Q}] = p-1$, from algebraic number theory we know that (p) can have at most $p-1$ factors, hence the previous decomposition of (p) is in fact a prime decomposition, so we also get that $(1-\zeta_p)$ is a prime ideal in $\mathcal{O}_K$. □

2.3 Preliminaries

We now move to definitions and facts needed specifically for Kummer's proof. The following propositions and lemmas are crucial in Kummer's proof. The following lemmas allow us to relate the algebraic integers in $\mathcal{O}_K$ with the rational integers in $\mathbb{Z}$. Kummer's main breakthrough in his proof was to work in an extension of $\mathbb{Q}$ where $(x^p + y^p)$ decomposed, so within the proof, he must go back and forth when dealing with elements of $\mathcal{O}_K$ and elements of $\mathbb{Z}$ using properties outlined by these lemmas.

Lemma 5. *Suppose $\alpha = a_0 + a_1\zeta_p + \cdots + a_{p-1}\zeta_p^{p-1}$ with each $a_i \in \mathbb{Z}$. If $a_i = 0$ for at least one i, then if $n \in \mathbb{Z}$ such that $n \mid \alpha$, then $n \mid a_j$ for all j.*

Proof. We know that $1 + \zeta_p + \cdots + \zeta_p^{p-1} = 0$, hence any $p-1$ elements of $\{1, \zeta_p, \ldots, \zeta_p^{p-1}\}$ is a basis for $\mathcal{O}_K$ over $\mathbb{Z}$. By assumption $a_i = 0$, so we choose the corresponding basis without ζ_p^i. The other coefficients make α an element of $\mathcal{O}_K$ with respect to this basis. Hence, if $n \mid \alpha$, then n must divide the coefficients of the basis representation of α, *i.e.* $n \mid a_j$ for each j. □

Lemma 6. *Let $\alpha \in \mathcal{O}_K$. Then α^p is congruent $\bmod p$ to an element of $\mathbb{Z}$.*

Proof. Take $\{1, \ldots, \zeta_p^{p-2}\}$ as the basis of $\mathcal{O}_K$. We can then write $\alpha = a_0 + a_1\zeta_p + \cdots + a_{p-2}\zeta_p^{p-2}$, where $a_i \in \mathbb{Z}$. This gives

$$\alpha^p \equiv a_0^p + (a_1\zeta_p)^p + \cdots + (a_{p-2}\zeta_p^{p-2})^p \equiv a_0^p + a_1^p + \cdots + a_{p-2}^p \bmod p,$$

since all nontrivial binomial coefficients are congruent to $0 \bmod p$. □

Lemma 7. *Assume x, y, z are a nontrivial solution to the equation $x^p + y^p = z^p$. The ideals $(x + \zeta_p^i y)$ with i ranging between $\{0, \ldots, p-1\}$ are either relatively prime as ideals or have exactly 1 common factor $(1-\zeta_p)$ such that the ideals generated by the quotients $\frac{x+\zeta_p^i y}{1-\zeta_p}$ are relatively prime.*

Proof. We make the assumption that x and y are relatively prime. Suppose $\exists\ \mathfrak{p}$ a prime ideal of $\mathcal{O}_K$ such that $\mathfrak{p} \mid (x + \zeta_p^i y)$ and $\mathfrak{p} \mid (x + \zeta_p^j y)$. From above, we know that $(1 - \zeta_p) = (1 - \zeta_p^k)$ as ideals when $p \nmid k$. Then $\mathfrak{p} \mid (x + \zeta_p^i y) - (x + \zeta_p^j y)$. However,

$$(x + \zeta_p^i y) - (x + \zeta_p^j y) = (\zeta_p^i y - \zeta_p^j y) = (1 - \zeta_p)(y).$$

Hence, $\mathfrak{p} \mid (1 - \zeta_p)$ or $\mathfrak{p} \mid (y)$. Similarly, we know that $(x + \zeta_p^i y) = (\zeta_p^{j-i} x + \zeta_p^j y)$, hence $\mathfrak{p} \mid (\zeta_p^{j-i} x + \zeta_p^j y) - (x + \zeta_p^j y)$. Since $(\zeta_p^{j-i} x - x) = (1 - \zeta_p^{j-i})(x) = (1 - \zeta_p)(x)$, we get that $\mathfrak{p} \mid (1 - \zeta_p)$ or $\mathfrak{p} \mid (y)$. Since x and y are coprime, one of these two statements implies that $\mathfrak{p} \mid (1 - \zeta_p)$. However, since $(1 - \zeta_p)$ is a prime ideal, we in fact get equality. Furthermore, note that if $(1 - \zeta_p) \mid (x + \zeta_p^k y)$, then $(1 - \zeta_p) \mid (x + \zeta_p^{k+1} y)$ since

$$(x + \zeta_p^{k+1} y) = (x + \zeta_p^k y) + (\zeta_p^k)(\zeta_p - 1)(y).$$

Thus, if $(1-\zeta_p)$ is a factor of $(x+\zeta_p^i y)$ for one i, then it is a factor for all i. In particular, we get that $x+y \equiv 0 \bmod 1 - \zeta_p$. Since $x+y \in \mathbb{Z}$, then $x+y \equiv 0 \bmod p$, however $x^p+y^p \equiv x+y \bmod p$, hence $z \equiv z^p \equiv 0 \bmod p$, *i.e.* $p \mid z$. If $p \nmid z$, we've arrived at a contradiction here, and thus $(1-\zeta_p)$ cannot be a common factor so in fact, the ideals $(x + \zeta_p^i y)$ have no common factors. If $p \mid z$, then we have that the only common factor between any two $(x + \zeta_p^i y)$ and $(x + \zeta_p^j y)$ is $1 - \zeta_p$.

It remains to be shown that $(1-\zeta_p)^2$ is not a factor of any two $(x+\zeta_p^i y)$ and $(x+\zeta_p^j y)$. Recall that we are assuming that $p \mid z$, hence we can further assume that $p \nmid y$; if this were the case, then $p \mid x$ as well, and we could reduce the counterexample $x^p + y^p = z^p$ by a factor of p^p. (During the proof, we use this argument to claim that the counterexample x, y, z is relatively prime.) Without loss of generality, assume that $i > j$ and note that

$$(x + \zeta_p^i y) - (x + \zeta_p^j y) = \zeta_p^i y - \zeta_p^j y = \zeta_p^j y(\zeta_p^{i-j} - 1).$$

From the fact that $(1-\zeta_p) = (1-\zeta_p^{i-j}) = (\zeta_p^{i-j}-1)$ as ideals, we have that $1-\zeta_p$ divides $\zeta_p^{i-j}-1$ exactly once. Furthermore, since $1-\zeta_p \mid y$ then $p \mid y$ (since $y \in \mathbb{Z}$) and $1-\zeta_p$ is relatively prime to ζ_p^j, we have that $1-\zeta_p \mid (x+\zeta_p^i y)-(x+\zeta_p^j y)$ but $(1-\zeta_p)^2 \nmid (x+\zeta_p^i y)-(x+\zeta_p^j y)$. Hence, we know that the quotients $\frac{x+\zeta_p^i y}{1-\zeta_p}$ are relatively prime. □

For any prime p, $\mathbb{Q}(\zeta_p)$ is automatically a subfield of $\mathbb{C}$ but not of $\mathbb{R}$. We can see that the automorphisms of $\mathrm{Gal}(K/\mathbb{Q})$ never send $\mathbb{Q}(\zeta_p)$ into $\mathbb{R}$. Furthermore, one of these automorphisms is the map of conjugation, sending $a+bi \mapsto a-bi$, its conjugate. Since the automorphisms have a group structure, we can pair each automorphism $\sigma \in \mathrm{Gal}(K/\mathbb{Q})$ with its **conjugate**, the unique automorphism described by composing σ with the map of conjugation. Note that this is equivalent in pairing elements of $(\mathbb{Z}/p\mathbb{Z})^\times$ with their additive inverse. However, there is a large subfield K^+ in $\mathbb{Q}(\zeta_p)$ which sits inside of $\mathbb{R}$, and properties of K and this subfield K^+ gives us information on the number of independent elements of each field and relates the corresponding rings of integers, $\mathcal{O}_K$ and $\mathcal{O}_{K^+}$ with the use of Dirichlet's Unit Theorem, stated below.

Theorem 8 (Dirichlet's Unit Theorem)**.** *For any field K over $\mathbb{Q}$ with r real embeddings and s conjugate pairs of complex embeddings, the unit group $\mathcal{O}_K^\times$ is finitely generated with rank equal to*

$$\mathrm{rank}(\mathcal{O}_K^\times) = r+s-1.$$

Proposition 9. *Fix some odd prime p, and let $K = \mathbb{Q}(\zeta_p)$. We have the following properties.*

(1) K is a totally complex field, i.e. $\exists$ 0 real embeddings and $\frac{p-1}{2}$ pairs of conjugate complex embeddings.
(2) The maximal totally real subfield of K is $K^+ = \mathbb{Q}(\zeta_p+\zeta_p^{-1})$, i.e. $K \cap \mathbb{R} = \mathbb{Q}(\zeta_p+\zeta_p^{-1})$. Furthermore, $\mathcal{O}_{K^+} = \mathbb{Z}[\zeta_p+\zeta_p^{-1}]$ and $[K:K^+] = 2$.
(3) K and K^+ have the same unit rank, hence the embedding of the corresponding unit groups $\mathcal{O}_{K^+}^\times \hookrightarrow \mathcal{O}_K^\times$ has finite index.

Proof. (1) Since all nontrivial p-th roots of unity are primitive, the automorphisms $\zeta_p \mapsto \zeta_p^k$ are embeddings into $\mathbb{C}$ that cannot be entirely contained in $\mathbb{R}$. Thus, there are no real embeddings and there are $p-1$ complex embeddings, hence $r=0$ and $s=\frac{p-1}{2}$.
(2) Geometrically, we can see that $\zeta_p+\zeta_p^{-1} \in \mathbb{R}$ as their imaginary coefficients are additive inverses, hence $\mathbb{Q}(\zeta_p+\zeta_p^{-1})$ is a subfield of K entirely contained in $\mathbb{R}$. Note that ζ_p is the root of a polynomial in $\mathbb{Q}(\zeta_p+\zeta_p^{-1})[x]$ defined as $f(x) = x^2-(\zeta_p+\zeta_p^{-1})x+1$. Since $f(x)$ is degree 2 and $x-\zeta_p$ is not a polynomial in $\mathbb{Q}(\zeta_p+\zeta_p^{-1})[x]$, $f(x)$ is automatically the minimal polynomial for ζ_p over $\mathbb{Q}(\zeta_p+\zeta_p^{-1})$, hence $[K:\mathbb{Q}(\zeta_p+\zeta_p^{-1})] = 2$. This additionally shows that $\mathbb{Q}(\zeta_p+\zeta_p^{-1})$ is the maximal real subfield in K since we have already seen that K is not totally real.
(3) By Dirichlet's Unit Theorem, we know that the rank of $\mathcal{O}_K^\times$ is a $r+s-1 = \frac{p-1}{2}-1$. Furthermore, as K^+ is totally real, the rank of $\mathcal{O}_{K^+}^\times$ is $[K^+:\mathbb{Q}]-1 = \frac{p-1}{2}-1$. □

Units in $\mathcal{O}_K^\times$ can be be easily described in terms of units in $\mathcal{O}_{K^+}^\times$ since the maximal real subfield is rather large in such a manner that the index of the unit groups is finite. We show in the following proposition that any unit of K can be decomposed into a product of p-th root of unity and a totally real unit in $\mathcal{O}_{K^+}^\times$.

Proposition 10. *For any $u \in \mathcal{O}_K^\times$, $\exists\, v \in \mathcal{O}_{K^+}^\times$ and an integer r such that $u = \zeta_p^r v$. It follows that the index of $\mathcal{O}_{K^+}^\times$ in $\mathcal{O}_K^\times$ is p.*

Sketch of proof. Consider some arbitrary unit $u \in \mathcal{O}_K^\times$ and let $\alpha = \frac{u}{\overline{u}}$ where $\overline{u}$ denotes the image of u under the map of conjugation. It follows that α is an algebraic integer and additionally, $|\alpha| = 1$. Furthermore, $|\sigma_k(\alpha)| = 1$ for each $\sigma_k \in \mathrm{Gal}(K/\mathbb{Q})$ since for all k, $\sigma_k(\overline{u}) = \overline{\sigma_k(u)}$. It is a fact used often in algebraic number theory that any algebraic integer whose Galois conjugates all have norm 1 must be a root of unity, so in particular, $\frac{u}{\overline{u}} = \pm\zeta_p^k$ for some k. It remains to show that $\alpha = +\zeta_p^k$. Assuming otherwise, we arrive at the contradiction that either 2 or $\overline{u}$ is contained in the

prime ideal generated by $(1-\zeta_p)$ from expressing both u and $\overline{u} \bmod 1-\zeta_p$. These two statements cannot be true based on a technical norm argument (not discussed here) in addition to the fact that $\overline{u}$ is a unit. Hence, we have $\alpha = \zeta_p^k$ for some k. From here, we find r such that $2r \equiv k \bmod p$, we set $v = \zeta_p^{-r}u$, hence $u = \zeta_p^r v$. (Note that if $\alpha = -\zeta_p^k$, then finding such an r does not work.) We see that $\overline{v} = \overline{\zeta_p^{-r}u} = \zeta_p^r\overline{u}$, hence $\frac{v}{\overline{v}} = \frac{\zeta_p^{-r}u}{\zeta_p^r\overline{u}} = \zeta_p^{-2r}\alpha = 1$, so v is in fact real, and therefore an element of K^+. □

Lemma 11 (Kummer's Lemma)**.** *If p is a regular prime and u is in $\mathcal{O}_K^\times$ such that u is congruent $\bmod p$ to an element of $\mathbb{Z}$, then u is a p-th power of an element of $\mathcal{O}_K^\times$.*

The statement above, although seemingly simple, uses a lot of machinery, including the class number formula, p-adic L-functions, and the characterization of regular primes using Bernoulli numbers. In fact, when Kummer first defined regular primes, he included this property as another condition and proved it much later. Kummer's proof of this statement is given in [Ed].

2.4 Case I: The Main Argument

We are now ready to present the proof when $p \nmid xyz$, generally known as the first case of Fermat's Last Theorem for regular primes. With this added assumption, Lemma 7 proves that the ideals $(x + \zeta_p^i y)$ are pairwise coprime. The main steps in this proof are obtained from this fact and the regularity of p, and are also used in the main argument for the second case and its generalizations in the following sections. The proof from this section uses the main ideas from Kummer's proof but is reformulated in the language of modern mathematics and uses some new lemmas. It is based on the proof in [Wa]. We restate the assumptions for this case of the theorem that will be proved in this section.

Theorem 12. *Suppose $p > 3$ is a regular prime. Then*

$$x^p + y^p = z^p, \quad p \nmid xyz$$

has no nontrivial solutions in the rational integers, i.e. any integer solution (x, y, z) has the property that $xyz = 0$.

Proof. Fix some regular prime $p > 3$, and assume that we have a nontrivial $x, y, z \in \mathbb{Z}$ satisfying the hypothesis. First, we can assume x, y, z are relatively prime. (Otherwise, we could divide by their greatest common denominator to get another counterexample.) Additionally, we show that for any such counterexample (x, y, z) we can rearrange to ensure that $x \not\equiv y \bmod p$ (which will be needed later). Suppose that $x \equiv y \equiv -z \bmod p$. Then note that

$$z \equiv z^p \equiv x^p + y^p \equiv x + y \equiv -2z \bmod p \quad \Longrightarrow \quad 3z \equiv 0 \bmod p,$$

then $p \nmid z$ implies $p \mid 3$, a contradiction to the fact that $p > 3$. Since we know that $x \equiv y \not\equiv -z \bmod p$, we can exchange y and $-z$ to get another counterexample satisfying all the hypotheses and $x \not\equiv y \bmod p$.

Kummer's Main Argument

In $\mathcal{O}_K$, we have the decomposition of ideals

$$(z)^p = (z^p) = (x^p + y^p) = (x + y)(x + \zeta_p y) \cdots (x + \zeta_p^{p-1} y),$$

and furthermore, all ideals on the right hand side are pairwise relatively prime. Since this decomposition is equal to the p-th power of the ideal generated by z, we have that each $(x + \zeta_p^i y)$ must be a p-th power of an ideal. (We can see this by considering the decomposition of (z) into prime ideals $\mathfrak{p}$. Since no $\mathfrak{p}$ is shared between various $(x + \zeta_p^i y)$, then in the corresponding decomposition of $(z)^p$, each $\mathfrak{p}^p$ is a factor of exactly one $(x + \zeta_p^i y)$.) Explicitly, we can write $(x + \zeta_p^i y) = \mathfrak{I}_i^p$ where $\mathfrak{I}_1\mathfrak{I}_2 \cdots \mathfrak{I}_{p-1} = (z)$, and each $\mathfrak{I}_i^p$ is principal. This is where we use the regularity of p, and

hence this is why the proof is limited to primes which do no divide the class number. Since each $\mathfrak{I}_i^p$ is principal, then in the class group defined as the group of ideals of $\mathcal{O}_K$ modulo the group of principal ideals of $\mathcal{O}_K$, we find that $\mathfrak{I}_i^p$ is trivial in the quotient group. However, the class group has order h_K which is not divisible by p, so there cannot exist a nontrivial element that has p-torsion, *i.e.* that is annihilated by the exponent p. Thus, $\mathfrak{I}_i$ must also be trivial in the class group, hence $\mathfrak{I}_i$ is also principal. Here, we see that if there was no assumption on the divisibility of h_K by p, then in fact $\mathfrak{I}_i$ need not be principal.

Since $\mathfrak{I}_i$ is principal, let $\alpha_i \in \mathcal{O}_K$ be its generator. Thus, $(x + \zeta_p^i y) = (\alpha_i)^p = (\alpha_i^p)$ hence $x + \zeta_p^i y = u\alpha_i^p$ for some unit $u \in \mathcal{O}_K^\times$. Note that we only need to treat the case for $i = 1$ (as we know that all nontrivial p-th roots of unity are primitive so based on the choice of ζ_p we can cycle through all cases). From Proposition 3.6, we can write $u = \zeta_p^r v$ where r is an integer and $v = \overline{v}$ is an element of $\mathcal{O}_{K^+}^\times$. By Lemma 6, $\exists$ a rational integer $a \in \mathbb{Z}$ such that $\alpha_i^p \equiv a \bmod p$. Thus, $x + \zeta_p y = \zeta_p^r v\alpha_i^p \equiv \zeta_p^r va \bmod p$. Furthermore, we get

$$x + \zeta_p^{p-1} y = x + \zeta_p^{-1} y = \zeta_p^{-r} v\overline{\alpha_i}^p \equiv \zeta_p^{-r} v\overline{a} \equiv \zeta_p^{-r} va \bmod p.$$

We know that $x+\zeta_p y \equiv \zeta_p^r va \equiv \zeta_p^{2r}(x+\zeta_p^{-1}y) \bmod p$ if and only if $x+\zeta_p y-\zeta_p^{2r}x-\zeta_p^{2r-1}y \equiv 0 \bmod p$. If 1, ζ_p, ζ_p^{2r}, ζ_p^{2r-1} are distinct, then by Lemma 5, $p \mid x, y$ a contradiction, and we are done. We know that 1 and ζ_p must be distinct, and similarly, ζ_p^{2r} and ζ_p^{2r-1} must also be distinct. Thus, we are left with 3 cases that each hinge on Lemma 5:

1. $1 = \zeta_p^{2r}$: From this, we get $x + \zeta_p y - x - \zeta_p^{-1} y \equiv 0 \bmod p$, *i.e.* $\zeta_p y - \zeta_p^{p-1} y \equiv 0 \bmod p$, hence from Lemma 5, $p \mid y$, a contradiction.

2. $\zeta_p = \zeta_p^{2r-1}$: This assumption reduces the congruence $x+\zeta_p y-\zeta_p^{2r}x-\zeta_p^{2r-1}y \equiv 0 \bmod p$ to $x - \zeta_p^2 x \equiv 0 \bmod p$. Hence again from Lemma 5, $p \mid x$, a contradiction.

3. $1 = \zeta_p^{2r-1}$: Note that this is equivalent to the relation $\zeta_p = \zeta_p^{2r}$, which reduces the congruence $x + \zeta_p y - \zeta_p^{2r} x - \zeta_p^{2r-1} y \equiv 0 \bmod p$ to $(x - y) - \zeta_p(x - y) \equiv 0 \bmod p$, so by Lemma 5, $p \mid x - y$, *i.e.* $x \equiv y \bmod p$, a contradiction to the choice of (x, y, z) made at the beginning of the proof. This proves that such a counterexample cannot exist, and the proof is complete. □

Kummer's original proof did not end in the same manner. After showing that $x+\zeta_p y = \zeta_p^r v\alpha$, Kummer found a congruence similar to $x + \zeta_p y - \zeta_p^{2r} x - \zeta_p^{2r-1} y \equiv 0 \bmod p$, and looked at coefficients using the binomial expansion of $1 + (\zeta_p - 1)^{r-1}$ to show that such an r cannot exist. However, the main argument Kummer was able to make was in showing that $\mathfrak{I}_i$ were principal, and thus $(x + \zeta_p^y)$ where p-th powers of algebraic integers in $\mathcal{O}_K$. The final case $p \mid z$ still rests upon this main property, and is in fact, a reduction to the proof of the first case of Fermat's Last Theorem for regular primes.

2.5 Case II: Completing the Proof

In this section, we finish the proof by assuming $p \mid z$. We can make this stronger assumption instead of $p \mid xyz$ since for any counterexample (x, y, z) we can assume x, y, and z, are pairwise coprime, so p only divides one of x, y, or z. We can rearrange and flip signs such that $p \mid z$. In this situation, Lemma 7 proves that the ideals $(x + \zeta_p^i y)$ have exactly one common factor, the prime ideal $(1 - \zeta_p)$. The proof from this section is the reformulation of Kummer's original proof for the second case in modern language. This proof uses the same main argument as the first case, but also involves the method of infinite descent in which a contradiction is reached by showing that if there is one "smallest" counterexample, then we can continue to construct "smaller" counterexamples ad infinitum. We restate the assumptions for this second case of the theorem that will be proved in this section.

Theorem 13. *Suppose $p > 3$ is a regular prime. Then*

$$x^p + y^p = z^p, \quad p \mid z$$

has no nontrivial solutions in the rational integers, i.e. any integer solution (x, y, z) has the property that $xyz = 0$.

Proof. We prove a stronger statement: There are no nontrivial solutions to $x^p + y^p = U(1 - \zeta_p)^{kp} z_0^p$ where $x, y, z_0 \in \mathcal{O}_K$ and $U \in \mathcal{O}_K^\times$ and relatively prime to each other as well as $1 - \zeta_p$. Note that the actual theorem is then just a special case where z is just written out as a product of its p-part and z_0, and x, y, z_0 are all integers.

Assume we have a counterexample satisfying the hypotheses. We have the decomposition of ideals $U(1 - \zeta_p)^{kp} z_0^p = (x + y)(x + \zeta_p y) \cdots (x + \zeta_p^{p-1} y)$. By this equality, we know that for some i, $1 - \zeta_p \mid x + \zeta_p^i y$, but by the same argument in Lemma 7, this implies that for all i, $x + \zeta_p^i y$ is divisible by $1 - \zeta_p$, and furthermore, the quotients $\frac{x+\zeta_p^i y}{1-\zeta_p}$ generate ideals which are pairwise relatively prime, again following from the same argument. We use the following lemma which allows us to assume that x and y are congruent to rational integers a and b modulo $(1 - \zeta_p)^2$.

Lemma 14. *For any algebraic integer $\alpha \in \mathcal{O}_K \backslash (1 - \zeta_p)$, $\exists\, l$ such that $\zeta_p^l \alpha \equiv a \bmod (1 - \zeta_p)^2$ where $a \in \mathbb{Z}$.*

Proof of Lemma. Note that $\mathcal{O}_K = \mathbb{Z}[\zeta_p] = \mathbb{Z}[1 - \zeta_p]$, hence one integral basis for $\mathcal{O}_K$ involves the powers of $(1 - \zeta_p)$. Therefore, we can find integers $a_0, a_1 \in \mathbb{Z}$ such that $\alpha \equiv a_0 + a_1(1 - \zeta_p) \bmod (1 - \zeta_p)^2$. Furthermore, since a_0 is nonzero outside of $(1 - \zeta_p)$, we can find $l \in \mathbb{Z}$ such that $a_1 \equiv a_0 l \bmod p$. Since $\zeta_p = 1 - (1 - \zeta_p)$, we have $\zeta_p^l \equiv 1 - l(1 - \zeta_p) \bmod (1 - \zeta_p)^2$. Thus,

$$\zeta_p^l \alpha \equiv (1 - l(1 - \zeta_p))(a_0 + a_1(1 - \zeta_p)) \equiv a_0 + (a_1 - l a_0)(1 - \zeta_p) \equiv a_0 \bmod (1 - \zeta_p)^2.$$

□

Returning the the proof of the theorem, we know that $\zeta_p^l x$ and $\zeta_p^j y$ are congruent to rational integers modulo $(1 - \zeta_p)^2$. Since we merely need x and y to satisfy the equation $U(1 - \zeta_p)^{kp} z_0^p = x^p + y^p$, exchanging them for $\zeta_p^l x$ and $\zeta_p^j y$ does not change anything. We know that $x + y \equiv a + b \bmod (1 - \zeta_p)^2$, where $a, b \in \mathbb{Z}$ are the integers congruent to x, y respectively. Since $1 - \zeta_p \mid x + y$, then $1 - \zeta_p \mid a + b$, which implies $p \mid a + b$ since $a + b \in \mathbb{Z}$. This, in turn, proves that $(1 - \zeta_p)^2 \mid x + y$ which tells us that k must be strictly greater than 1. To use the method of infinite descent, we choose our nontrivial counterexample (x, y, z_0) such that k is minimal. Our contradiction will arise from the construction of a new counterexample (x', y', z_0') such that $x'^p + y'^p = U'(1 - \zeta_p)^{(k-1)p} z_0'^p$.

From above, we know that $(1 - \zeta_p)^2 \mid x + y$, and by Lemma 7, we know that the quotients $\frac{x+\zeta_p^i y}{1-\zeta_p}$ are relatively prime. Hence all of the extra powers of $(1 - \zeta_p)$ divide $x + y$ only. Since $(1 - \zeta_p)^{p-1} \mid (x + \zeta_p y)(x + \zeta_p^2 y) \cdots (x + \zeta_p^{p-1} y)$ exactly, (*i.e.* $(1 - \zeta_p)^p \nmid (x + \zeta_p y)(x + \zeta_p^2 y) \cdots (x + \zeta_p^{p-1} y)$). We know further that $(1 - \zeta_p)^{kp} \mid x^p + y^p$ exactly, hence $(1 - \zeta_p)^{kp-p-1} \mid x + y$. Thus, $(1 - \zeta_p)^{(k-1)p} \mid \frac{x+y}{1-\zeta_p}$ exactly. This will be crucial when we consider the ideal generated by the quotients.

Changing Fermat's equation to ideals, we have

$$\left((1 - \zeta_p)^{k-1} z_0\right)^p = \prod_{i=0}^{p-1} \left(\frac{x + \zeta_p^i y}{1 - \zeta_p}\right),$$

where the ideals on the right are relatively prime. As in the first case, by Kummer's main argument, we have that each ideal generated by $\frac{x+\zeta_p^i y}{1-\zeta_p}$ is a p-th power of a principal ideal, hence $\exists\, \alpha_i \in \mathcal{O}_K$

such that we have the equalities $\frac{x+\zeta_p^i y}{1-\zeta_p} = u_i \alpha_i^p$ where u_i are units in $\mathcal{O}_K^\times$. Furthermore, we know that $\{\alpha_0, \ldots, \alpha_{p-1}\}$ are pairwise relatively prime since their p-th powers are relatively prime. From the previous argument, we know that $(1-\zeta_p)^{k-1} \mid \alpha_0$, and we can furthermore write $\alpha_0 = (1-\zeta_p)^{k-1}\beta$ where β is relatively prime to $1-\zeta_p$. From the equalities of $x+\zeta_p^i y$, we use $x+\zeta_p y$ and $x+\zeta_p^{p-1}y = x+\zeta_p^{-1}y$ as well as the new substitution for $x+y$ to get

$$(1-\zeta_p)^{(k-1)p} u_0 \beta^p - u_1 \alpha_1^p = \frac{(x+y)-(x+\zeta_p y)}{1-\zeta_p} = y$$

$$\zeta_p (1-\zeta_p)^{(k-1)p} u_0 \beta^p - \zeta_p u_{-1} \alpha_{-1}^p = \frac{(x+\zeta_p^{-1}y)-(x+y)}{\zeta_p^{-1}(1-\zeta_p)} = y$$

$$\left[\zeta_p (1-\zeta_p)^{(k-1)p} u_0 \beta^p - \zeta_p u_{-1} \alpha_{-1}^p\right] - \left[(1-\zeta_p)^{(k-1)p} u_0 \beta^p - u_1 \alpha_1^p\right] = 0$$

If we let $U' := \frac{(1+\zeta_p)u_0}{-u_1}$ and $V' := \frac{\zeta_p u_{-1}}{-u_1}$, then U' and V' are units and the last equation simplifies to $U'(1-\zeta_p)^{(k-1)p}\beta^p = \alpha_1^p + V'\alpha_{-1}^p$. If we consider the equation modulo p, then since $p \mid (1-\zeta_p)^{p-1}$, we have $0 \equiv \alpha_1^p + V'\alpha_{-1}^p \bmod p$. Recall from Lemma 6, we know that $\alpha_{\{1,-1\}}^p \equiv a_{\{1,-1\}} \bmod p$ where $a_1, a_{-1} \in \mathbb{Z}$, thus $0 \equiv a_1 + V' a_{-1} \bmod p$. (Note that a_1 and a_{-1} are nonzero as they are relatively prime to α_0 which is divisible by p.) However, this implies that V' must in fact be congruent to a rational integer$\bmod p$. Kummer's Lemma then allows us to rewrite V' as a p-th power of some unit $v \in \mathcal{O}_K^\times$. If we let $x' := \alpha_1$, $y' := v\alpha_{-1}$, and $z_0' = \beta$, then we have $U'(1-\zeta_p)^{(k-1)p} z_0'^p = x'^p + y'^p$, another counterexample which contradicts the minimality of k. This completes the proof for the second case, and thus Fermat's Last Theorem holds for regular primes. □

2.6 A Modern View

Kummer's proof of Fermat's Last Theorem can be reformulated to involve a modern approach that was attempted for the proof of the entire theorem. For any counterexample at prime p, the goal is to attach a representation ρ over $K = \mathbb{Q}(\zeta_p)$ from the algebraic closure $\overline{\mathbb{Q}(\zeta_p)}$ into a certain extension L viewed as vector spaces over K. In particular, the extension L/K would be equipped with Galois group isomorphic to $(\mathbb{Z}/p\mathbb{Z})^\times$. Note that this gives rise to a map $\mathrm{Gal}(\overline{\mathbb{Q}(\zeta_p)}/\mathbb{Q}(\zeta_p)) \to \mathrm{Gal}(L/K) \hookrightarrow \mathbb{F}_p$. We do this by considering a different interpretation of regularity involving class field theory. Global class field theory tells us that there exists an extension of $K = \mathbb{Q}(\zeta_p)$ known as the **Hilbert class field** H_K with the defining property that $\mathrm{Gal}(H_K/K)$ is isomorphic to the ideal class group of K. Furthermore, we know that H_K is **totally unramified**, *i.e.* none of the prime ideals in K have a decomposition in H_K with repeated factors. Galois theory explains the extensions that lie between K and H_K in relation to $\mathrm{Gal}(H_K/K)$, *i.e.* in relation to the structure of the ideal class group of K. For example, if p does not divide h_K, there is no p-torsion in the ideal class group, *i.e.* there is definitely no extension of K of degree p that is in H_K. In particular, there does not exist a cyclic extension of degree p which is totally unramified. Hence, to every counterexample at a prime p, we construct a totally unramified extension L over $\mathbb{Q}(\zeta_p)$ with degree p in order to come to a contradiction. In terms of representations, we note that $\mathrm{Gal}(L/\mathbb{Q}(\zeta_p)) \cong (\mathbb{Z}/p\mathbb{Z})^\times$, so by adding the zero automorphism we get $\mathbb{F}_p$. Facts from infinite Galois theory and Kummer theory allow us to form the representation ρ from the full Galois group of $\mathbb{Q}(\zeta_p)$ to $\mathbb{F}_p$. As expected, such ρ do not exist at regular primes p. The following proof is based on [Pa]. For this proof, we need some more facts, the most important of which come from class field theory. As usual, let $K = \mathbb{Q}(\zeta_p)$, and we take p to be a prime greater than 3.

Proposition 15. *(1) If u is a unit of $\mathcal{O}_K$ such that it is congruent to a rational integer modulo p and not a p-th power in $\mathcal{O}_K$, then the field extension $K(u^{1/p})/K$ is a cyclic extension of order p that has the property of being totally unramified. (This holds for any p-th root of u.)*
(2) If $\mathfrak{I}$ is an ideal of $\mathcal{O}_K$ such that $\mathfrak{I}^p$ is principal, but $\mathfrak{I}$ is not, then there is a cyclic extension over K of order p that has the property of being totally unramified.

It is easy to see that the hypotheses in the above proposition never hold for primes which are regular (this follows from the definition and Kummer's Lemma), hence one can understand that at such regular primes, a cyclic extension would never exist.

Theorem 16. *If there exists $x, y, z \in \mathbb{Z}$ such that $x^p + y^p = z^p$, then we can produce a cyclic extension L of $K = \mathbb{Q}(\zeta_p)$ of order p which has the property of being totally unramified.*

Proof. As expected, we have two cases, $p \nmid xyz$ and $p \mid z$. Furthermore, we assume as usual that x, y, and z are relatively prime. In this proof, we will reference the original proof from the previous sections, using the exact same notation.

Case 1. Assume $p \nmid xyz$. As in the original proof, we rearrange and flip signs such that $x \not\equiv y \bmod p$. We use the usual decomposition into relatively prime ideals $(z)^p = (x+y)(x+\zeta_p y)\cdots(x+\zeta_p^{p-1}y)$ so that we can write $(x+\zeta_p^i) = \mathfrak{I}_i^p$ for all $i \in \{0, \ldots, p-1\}$. Here, we don't have the assumption that p is regular. However, from the argument of the original proof, we know that if $\mathfrak{I}$ is principal, then following the same argument, we obtain the congruence $x+\zeta_p y-\zeta_p^{2r}x-\zeta_p^{2r-1}y \equiv 0 \bmod p$ and eventually get contradictions to $p \nmid xyz$ or $x \not\equiv y \bmod p$ using Lemma 7. Thus, if x, y, and z exist, it must be that $\mathfrak{I}$ must be principal, hence by Proposition 6.1, there exists an extension of K with the needed properties.

Case 2. Assume $p \mid z$. Here, we generalize and consider the equation $x^p + y^p = U(1-\zeta_p)^{kp} z_0^p$ where x, y, z_0 are elements of $\mathcal{O}_K$ such that they are relatively prime to each other as well as $1-\zeta_p$, and we consider a solution such that k is minimal. We note that in the decomposition $U(1-\zeta_p)^{kp}z_0^p = (x+y)(x+\zeta_p y)\cdots(x+\zeta_p^{p-1}y)$, each of factors $x+\zeta_p^{p-1}y$ is divisible by $1-\zeta_p$ and we can change x and y accordingly such that $(1-\zeta_p)^{k(p-1)+1} \mid x+y$. It follows that $\left((1-\zeta_p)^{k-1}z_0\right)^p = \prod_{i=0}^{p-1}\left(\frac{x+\zeta_p^i y}{1-\zeta_p}\right)$, and the ideals generated by $\frac{x+\zeta_p^i y}{1-\zeta_p}$ are p-th power of ideals $\mathfrak{I}_i \in \mathcal{O}_K$. Here since we do not have the assumption that p is regular, we do not know whether or not these ideals are principal. Nevertheless, if these ideals are principal, in particular, if $\mathfrak{I}_0$, $\mathfrak{I}_1$, and $\mathfrak{I}_{-1} = \mathfrak{I}_{p-1}$ are principal, then we have the following three equations

$$\begin{aligned} x+y &= (1-\zeta_p)^{kp+1}u_0\beta^p \\ x+\zeta_p y &= (1-\zeta_p)u_1\alpha_1^p \\ x+\zeta_p^{-1}y &= (1-\zeta_p)u_{-1}\alpha_{-1}^p \end{aligned}$$

where $\beta, \alpha_1, \alpha_{-1}$ are elements of $\mathcal{O}_K$ and u_0, u_1, u_{-1} are units in $\mathcal{O}_K^\times$. Following the same argument, we arrive at an equation $U'(1-\zeta_p)^{(k-1)p}\beta^p = \alpha_1^p + V\alpha_{-1}^p$ where U' and V' are units of K. Looking at this equation modulo p, we arrive at the conclusion that V' is congruent to an integer modulo p, so we either have the case that V' is not a p-th power producing a cyclic extension $L = K(V'^{1/p})$ from Proposition 6.1 satisfying all needed properties or there exists $v \in \mathcal{O}_K^\times$ such that $V' = v^p$. If we let $x' := \alpha_1$, $y' := v\alpha_{-1}$, and $z_0' := \beta$, then we have $U'(1-\zeta_p)^{(k-1)p}z_0'^p = x'^p + y'^p$, a contradiction to the minimality of k. Hence, we see that one of $\mathfrak{I}_1$, $\mathfrak{I}_0$, or $\mathfrak{I}_{-1}$ must not be principal, and by Proposition 6.1, we can produce the extension L over K with the needed properties. □

From the existence of such an L, we can go further to produce a representation $\rho : \mathrm{Gal}(\overline{K}/K) \to \mathrm{Gal}(L/K) \cong \mathbb{F}_p$. From Galois theory, we know that $\mathrm{Gal}(\overline{K}/K)$ can be expressed as an inverse limit of the Galois groups of its finite Galois subextensions, including L. In particular, we have a canonical homomorphism $\mathrm{Gal}(\overline{K}/K) \to \mathrm{Gal}(\overline{K}/K)/\mathrm{Gal}(\overline{K}/L) \cong \mathrm{Gal}(L/K)$, giving rise to the exact representation ρ that we need.

For the complete the proof for Fermat's Last Theorem, Andrew Wiles attempted to associate to every counterexample (x, y, z, p), a representation $\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_n(\mathbb{F}_p)$ such that the representation was unramified away from p and had "nice" ramification at p. Wiles immediately followed by proving no such representations exist, contradicting the existence of such counterexamples ([Pa]). This is comparable to the strategy used here to prove Fermat's Last Theorem for regular primes.

2.7 Generalizing the Second Case

The method used in the second case of Kummer's proof can be generalized to prove Fermat's Last Theorem for more than just regular primes. The basic argument stays the same, but we instead consider the generalized equation

$$x^p + y^p = u\lambda^m z^p,$$

where p is an odd prime, $\lambda = (1-\zeta_p)^2$, $x, y, z, \in \mathbb{Z}[\lambda]$ such that they are relatively prime with each other and λ, $u \in \mathbb{Z}[\lambda] \cap \mathbb{R}$, and $m \geq \frac{p(p-1)}{2}$. We want to show that this equation has no solutions. We must make two assumptions on our choice of p that loosely generalize the property of regularity. The first is that $p \nmid h_{K^+}$, *i.e.* p must not divide the class number of $K^+ = \mathbb{Q}(\zeta_p + \zeta_p^{-1})$. The second assumption is that certain units η arising within the argument can be expressed as a p-th power of a unit in K^+. We will state this assumption more accurately when these units show up. For this argument, we will need a couple of facts. The proof for the following lemma can be found in [Wa].

Lemma 17. *(1) For any $\alpha \in \mathcal{O}_K$ such that $\alpha \equiv 1 \bmod (1-\zeta_p)^p$, the extension $K(\alpha^{1/p})/K$ is unramified at $(1-\zeta_p)$.*
(2) Assume $p \nmid h_{K^+}$. Let $\alpha \in \mathcal{O}_K$ be such that $\overline{\alpha} = \alpha^{-1}$ and $K(\alpha^{1/p})/K$ is unramified. Then there exists $\beta \in K$ such that $\alpha = \beta^p$.

Main Argument. We are ready for the main argument. We will only present a sketch of the proof. The full detailed proof can be found in [Wa].

Assume there exists a solution to the equation $u\lambda^m z^p = x^p + y^p$ satisfying the properties described above. We can decompose the right hand side in K to get $u\lambda^m z^p = (x+y)(x+\zeta_p y)\cdots(x+\zeta_p^{p-1}y)$. The only common factor of any two $(x+\zeta_p^i y)$ and $(x+\zeta_p^j y)$ for $i \neq j$ is the prime ideal $(1-\zeta_p)$. Furthermore, we know that $(1-\zeta_p)^2 = (\lambda) \mid (x+y)$, so we have the equality

$$(x+y)\left(\frac{x+\zeta_y}{1-\zeta_p}\right)\left(\frac{x+\zeta_y^2}{1-\zeta_p}\right)\cdots\left(\frac{x+\zeta_y^{p-1}}{1-\zeta_p}\right) = v\lambda^{m-(p-1)/2}z^p,$$

where v is a unit and the algebraic integers on the right generate ideals which are pairwise relative prime, so in particular, since $1-\zeta_p \mid x+y$, $1-\zeta_p \nmid \frac{x+\zeta_p^i y}{1-\zeta_p}$ for any $i \in \{1,\ldots,p-1\}$. It follows that there exists ideals $\mathfrak{I}_i$ such that $\mathfrak{I}_0^p(\lambda)^{m-(p-1)/2} = (x+y)$ and $\mathfrak{I}_i^p = \frac{x+\zeta_p^i y}{1-\zeta_p}$ for all other i. Note that $\mathfrak{I}_{p-i} = \mathfrak{I}_{-i}$ is the complex conjugate of $\mathfrak{I}_i$.

If we assume that $p \nmid h_{K^+}$, then $\mathfrak{I}_0$ is principal in $\mathbb{Z}[\lambda]$. (Note that it is okay to think of $\mathfrak{I}_0$ in $\mathbb{Z}[\lambda]$ since $(1-\zeta_p) \nmid \mathfrak{I}_0$.) Furthermore, since $x+y$ and λ are elements of $\mathbb{R}$, the generator α_0 of $\mathfrak{I}_0$ is also real, so we get $x+y = u_0\lambda^{m-(p-1)/2}\alpha_0^p$ where u_0 is a unit which is also real. For any $i \neq 0$, define

$$a_i = -\zeta_p^{-i}\frac{x+\zeta_p^i y}{x+\zeta_p^{-i}y} \equiv 1 \bmod (1-\zeta_p)^{2m-p},$$

so in particular $a_i \equiv 1 \bmod (1-\zeta_p)^p$. Note that the principal ideal generated by a_i can be decomposed as a p-th power of $(\mathfrak{I}_i/\mathfrak{I}_{-i})$. Thus, from Lemma 7.1, we not only know that the extension $K(a_i^{1/p})/K$ is unramified at $(1-\zeta_p)$, we furthermore know that it is totally unramified. Additionally, from the second part of Lemma 7.1, $a_i = \beta_i^p$ where $\beta_i \in K$. This allows for us to find $\alpha_i \in \mathbb{Z}[\lambda]$ such that $\frac{x+\zeta_p^i y}{1-\zeta_p} = u_i\alpha_i^p$ where u_i is a real unit. Note that $(\overline{\alpha_i})^p = \alpha_{-i}^p$, so up to a root of unity, $\overline{\alpha_i} = \alpha_{-i}$.

From the equalities $x+\zeta_p^i y = (1-\zeta_p)u_i\alpha_i^p$ and $x+\zeta_p^{-i}y = (1-\zeta_p^{-i})u_{-i}\overline{\alpha_i}^p$ as well as the formula for $x+y$, we get

$$-xy = u_i^2(\alpha_i\overline{\alpha_i}) - u_0^2\lambda^{2m-p+1}\alpha_0^{2p}\lambda_a^{-1}.$$

For j such that $j \neq 0$ and $i \not\equiv \pm j \bmod p$, a similar equality holds. Combining the two gives us

$$\eta^2(\alpha_i\overline{\alpha_i})^p + (-\alpha_b\overline{\alpha_b})^p = \delta\lambda^{2m-p}(\alpha_0^2)^p,$$

where $\eta = u_i/u_j$ and δ is a real unit. Note that η defined here is the one needed as a p-th power of a unit from K^+ in the second assumption. This allows us to define $x_1 = \eta^{2/p}\alpha_a\overline{\alpha_a}$, $y_1 = -\alpha_b\overline{\alpha_b}$, and $z_1 = \alpha_0^2$ so that the above equations turn into $x_1^p + y_1^p = \delta\lambda^{2m-p}z_1^p$. It is easy to show that x_1, y_1, z_1 are pairwise relatively prime with λ. Again, we can use the method of infinite descent to produce a contradiction. If we assume that (z) has the smallest number of distinct prime ideal factors in its decomposition, we in fact know that $(z) = \mathfrak{I}_0\mathfrak{I}_1\cdots\mathfrak{I}_{p-1}$ where $\mathfrak{I}_i$ and $\mathfrak{I}_j$ are relatively prime for $i \neq j$. However, $z_1 = \alpha_0^2$ and α_0 is the generator of $\mathfrak{I}_0$, hence $(z_1) = \mathfrak{I}_0^2$ so $\mathfrak{I}_1, \ldots, \mathfrak{I}_{p-1}$ must be trivial. However, this implies that each $\frac{x+\zeta_p^i y}{1-\zeta_p}$ is a unit for $i \neq 0$. With some manipulation, we arrive at the fact that either $x+y = 0$ or $\zeta_p^2 = 1$, both contradictions. Altogether, we see that such a solution cannot exist. □

The above argument proves the second case of Fermat's Last Theorem for regular primes as well as other cases, although it remains to show why the two assumptions are satisfied by the property of regularity. Proofs demonstrating how to go about satisfying the two assumptions for regular primes as well as other cases can be found in [Wa].

References

[Ed] H. M. Edwards: *Fermat's Last Theorem*, 1st ed. New York: Springer-Verlag 1977.

[Pa] K. Paranjape: On (Kummer's Approach to) Fermat's Last Theorem, *Bona Math.* (1994).

[Ri] P. Ribenboim: *13 Lectures on Fermat's Last Theorem*, 1st ed. New York: Springer-Verlag 1979.

[vdP] A. van der Poorten: *Notes on Fermat's Last Theorem*, 1st ed. New York: Wiley-Interscience 1996.

[Wa] L. Washington: *Introduction to Cyclotomic Fields*, 2nd ed. New York: Springer-Verlag 1997.

STUDENT ARTICLE

3 Securing Your Hair

Grant Dasher[†]
Harvard University '09
Cambridge, MA 02138
gdasher@fas.harvard.edu

Abstract

Inspired by geometric intuition, the Braid Group admits a neat algebraic structure with complexity sufficient to suggest cryptographic applications. We investigate these applications and present an algorithm for normalizing words in the braid group. Such normalization is critical to a realistic implementation of a computational system based on braids. We also present a simple cryptographic protocol based on the braid group.

3.1 Introduction

The problem of transmitting a secure message over an insecure channel is nearly as old as time itself. However, it has taken on new significance in the last 50 years as the modern global financial infrastructure depends heavily on secure communication. For generations, the approach to secure communication remained basically unchanged: prearrange the transfer of some shared piece of information (the **secret**) and use it to encrypt a message. Although the algorithms used for encrypting the message changed with time, the technique itself did not.

In the 1970s, James Ellis, Clifford Cocks, and Malcolm Williamson discovered the beginnings of what came to be called **public-key** or asymmetric cryptography[El]. Asymmetric cryptography no longer required the pre-exchange of a secret piece of information. Instead, a piece of public information could be used to encrypt a message that could only be decrypted by a different, private, piece of information. The public-key technique that Ellis, Cocks, and Williamson discovered relied on the difficulty of factoring large prime numbers. Although this problem is still considered difficult in classical terms, there exist quantum algorithms which can solve the problem much faster (in fact, in polynomial time)[NC].

Although no one has yet built a useful quantum computer, the possibility worries most security theorists. As a result, there has been a great deal of research into other, potentially more secure, public-key cryptography systems. One such system, which drew a great deal of initial interest, relies on the Artin Braid Group. Although later research has revealed additional structure in the braid group which may limit its cryptographic potential, there still remain potentially interesting cryptographic applications of the group[De]. Regardless, the mathematics necessary to successfully implement a braid based cryptographic system is interesting in its own right.

In this paper, we describe the braid group geometrically and algebraically, discuss in detail the mathematics required to implement the braid group algebraically (specifically, we discuss the word problem and its solution: left weighted canonical form), and conclude by describing a simple cryptographic system which relies on braids.

3.2 The Braid Group

The Artin Braid Group is an infinite non-commutative group which in some sense generalizes the classical symmetric group by encoding additional topological information with each permutation. The group can be defined both algebraically and geometrically. The algebraic definition is most

[†]Grant Dasher '09 is a senior Mathematics and Computer Science concentrator at Harvard. He is interested in using mathematics to reason about the correctness of software and to design new programming languages.

useful for computational purposes, but we begin with the geometric definition because it best motivates the intuition behind the key structure theorems.

Definition 1. Let $\sigma \in S_n$ be a permutation. An ***n*-stranded braid** is a set of n non-intersecting smooth paths γ_i such that:

- $\gamma_i(0) = (i, 0, 1)$ and $\gamma_i(1) = (\sigma(i), 0, 0)$.
- For each $j \in [0, 1]$, each path intersects each plane $z = j$ exactly once.

It is natural to consider equivalence classes of braids. We say that two braids A_0 and A_1 are **equal** if they are isotopic as braids, that is, if there exists a continuous family of braids $\{A_t\}_{t\in(0,1)}$ carrying A_0 into A_1.

The set of equivalence classes of n-stranded braids, with the natural operation of concatenation (and rescaling), forms a group. It is clear that, up to isotopy, this operation is associative and invertible. It is also clear that the identity braid is the braid represented by n vertical parallel lines. We will refer to the braid group on n strands as B_n. We denote the identity element of this group by e. It is the trivial braid consistent of parallel straight lines.

The Braid Group is generated by the $n - 1$ elementary braids shown in Figure 3.1.

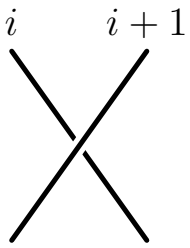


Figure 3.1: The generators σ_i

It is clear that after a bit of isotopy, any braid can be written as a word in these generators (and their inverses). Furthermore, a little thought will reveal these generators satisfy the following relations.

1. $\sigma_i\sigma_j = \sigma_j\sigma_i$, for $|j - i| \geq 2$ (far commutivity)
2. $\sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1}$ (braid relation)

It turns out that these two relations are the only relations satisfied by elements of the braid group (in addition to the trivial relations implied by the group axioms). A combinatorial proof of this can be found in [Ma]. A much more elegant (but more sophisticated) proof is in [Ha]. As a result, we have the following

Theorem 2. *B_m admits the following presentation:*

$$\langle \sigma_1, \ldots \sigma_{n-1} \,|\, \sigma_i\sigma_j = \sigma_j\sigma_i \textit{ for } |i - j| \geq 2 \,;\, \sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1} \rangle$$

3.3 Algebraic Implementation and Braid Normal Form

In order to implement braids on a computer, we need a representation. A standard representation would be an algebraic braid word, that is a formal list of symbols σ_i. Unfortunately, braid words are not unique. For example, in B_4, the words $\sigma_1\sigma_3\sigma_2\sigma_1$ and $\sigma_3\sigma_2\sigma_1\sigma_2$ are equivalent (it is clear that a simple application of the relations will transform one into the other). The problem of determining whether two words represent the same group element is known to group theorists as the word problem. In general, the problem is undecidable for an arbitrary group. However, there is a viable solution in the braid group. To solve the word problem, we must select a distinguished word from each equivalence class (called the **normal form** or **canonical form** of the braid). If we

have an algorithm for transforming an arbitrary word into its normal representative, this yields a solution to the word problem. Luckily, in the case of the braid group, a normal form and associated algorithm do exist.

In order to define braid normal form, we must first more carefully study the structure of the braid group. We limit our focus to the semi-group B_n^+ of positive braids, that is braids where only positive powers of the generators appear in any braid word. We shall see that this suffices, for we can uniquely express all braid words in terms of positive braid words. It is true, although we shall not prove it, that B_n^+ can be considered as an abstract semi-group on exactly the same generators and relations which define B_n and that this semi-group embeds into B_n. The proof of this fact basically amounts to showing B_n^+ is left cancelable, right cancelable, and right reversible. A full proof can be found in [Bi]. Various normal forms can be defined for the braid group, all of which employ the same basic ideas. We will closely follow the **left-weighed canonical form** defined in [EM].

There is a natural homomorphism $\mathrm{wt} : B_n \to \mathbb{Z}$ defined by $\mathrm{wt}(\sigma_i) = 1$. When restricted to B_n^+, this homomorphism coincides with the length of the braid word.

We can also define an *anti*-automoprhism $\mathrm{rev} : B_n \to B_n$ by $\mathrm{rev}(\sigma_i) = \sigma_i$. That is, rev is the unique map such that $\mathrm{rev}(\sigma_i) = \sigma_i$ and $\mathrm{rev}(AB) = \mathrm{rev}(B)\,\mathrm{rev}(A)$ for all $A, B \in B_n$. We can interpret this map geometrically as reading the braid word 'bottom up' instead of 'top down' as one normally would.

There is also a natural automorphism $\tau : B_n \to B_n$ given by $\tau(\sigma_i) = \sigma_{n-i}$ which can be realized geometrically as 'turning over' a braid. τ is an involution and in fact can be realized explicitly as conjugation by a particular braid.

Definition 3. The **fundamental braid** Δ_n is defined inductively by:

$$\Delta_n = \Delta_{n-1}\sigma_{n-1}\sigma_{n-2}\ldots\sigma_1$$

and $\Delta_1 = \sigma_1$.

Geometrically, we can interpret this braid as the braid that 'folds' the strings over themselves. Figure 3.2 shows an example in the case of $n = 4$. When there is no danger of confusion, we will write Δ for Δ_n.

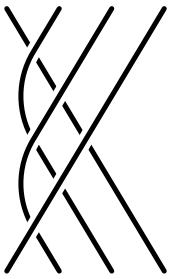


Figure 3.2: The fundamental braid Δ_4

Proposition 4. *Let $A \in B_n$ be a braid. Then $\tau(A) = \Delta_n A \Delta_n^{-1}$.*

Proof. It suffices to show that $\Delta_n\sigma_i = \sigma_{n-i}\Delta_n$. We omit the details, but this follows from a straightforward application of the relations and the inductive definition of Δ_n. □

We introduce a partial order $\leq$ on B_n as follows:

Remark. For $A, B \in B_n$, say $A \leq B$ if there exists $C_1, C_2 \in B_n^+$ such that $B = C_1 A C_2$.

It is clear that $B \in B_n^+ \iff e \leq B$ and $A \leq B \iff B^{-1} \leq A^{-1}$. It is also clear from the definition of Δ that each generator satisfies $e \leq \sigma_i \leq \Delta$. The partial order notation greatly simplifies the statement of the key algebraic relationships which lead to braid normal form. The key algebraic lemma is the following.

Lemma 5. *If $A \leq \Delta^s$, then $\Delta^s = D_1 A = A D_2$ for some $D_1, D_2 \in B_n^+$. Similarly, if $\Delta^r \leq A$, then $A = E_1 \Delta^r = \Delta^r E_2$ for some $E_1, E_2 \in B_n^+$.*

Proof. Since τ is simply conjugation by Δ, it is clear that $\Delta^s C_2 = \tau^s(C_2)\Delta^s$. To get the first statement from this fact, write $\Delta^s = C_1 A C_2$ with $C_1, C_2 \in B_n^+$. Then $\Delta^s = C_1 \tau^s(C_2) A$, and similarly, $\Delta^s = A\tau^s(C_1)C_2$ since $\tau^2 = 1$. To get the second statement, write $A = C_1 \Delta^r C_2$. We have $\tau^r(A) = \tau^r(C_1 \Delta^r C_2) = \tau^r(C_1)\tau^r(\Delta^r)\tau^r(C_2) = \tau^r(C_1)\Delta^r \tau^r(C_2) = \Delta^r C_1 \tau^r(c_2)$. The other equality is similar. □

As corollaries, we find that:

Corollary 6. *If $\Delta^{r_1} \leq B \leq \Delta^{s_1}$ and $\Delta^{r_1} \leq B \leq \Delta^{s_2}$, then*

$$\Delta^{r_1+s_1} \leq BC \leq \Delta^{s_1+s_2}.$$

Corollary 7. *For each braid B, there exist $r, s \in \mathbb{Z}$ such that $\Delta^r \leq B \leq \Delta^s$.*

Proof. Write B as a word in the generators $\sigma_i^{\pm 1}$ and use Corollary 6 together with fact that $e \leq \sigma_i \leq \Delta$ and $\Delta^{-1} \leq \sigma_i^{-1} \leq e$. □

Remark. For $A \in B_n$, we say that $A \in [r, s]$ if $\Delta^r \leq A \leq \Delta^s$.

We can now begin to construct the braid normal form. Before we state the main theorem, we need one more definition.

Definition 8. The **starting set**, $S(P) \subset \{1, \ldots, n-1\}$, of a positive braid P is the set

$$S(P) = \{i \mid P = \sigma_i P_i, P_i \geq e\}.$$

Similarly, the **finishing set** is

$$F(P) = \{i \mid P = P_i \sigma_i, P_i \geq e\}.$$

We can now state the main theorem, which we will prove over the remainder of this section.

Theorem 9 (Main Theorem)**.** *Let P be a positive braid. Then there is a unique expression $P = A_1 A_2 \ldots A_k$ with $A_i \in [0, 1]$, $A_k \neq e$, and $S(A_{i+1}) \subset F(A_i)$ for each i.*

It is also clear now why it suffices to consider positive braids. For, given an arbitrary braid B, we can find a maximal r such that $\Delta^r \leq B$ and hence write $B = \Delta^r P$, where P is a positive braid. If we can find a unique normal form for P, then the unique normal form for B follows simply as the power r together with the normal form for P.

Definition 10. A positive factorization $P = AB$, with $A, B \geq e$ is called **left-weighted** if $S(B) \subset F(A)$. Similarly, it is called **right-weighted** if $S(B) \supset F(A)$.

We will develop the theory of left-weighted factorizations, although one can equally well develop a theory of right-weighted factorizations with the same basic results. The key to proving the main theorem is the following

Lemma 11 (Main Lemma)**.** *Every positive braid P has a unique left-weighted factorization $P = A_1 P_1$ with $A_1 \in [0, 1]$. Such a factorization is universal in the sense that every other positive factorization $P = AB$, with $A \in [0, 1]$, satisfies $A_1 = AQ$ for some positive braid Q.*

In order to prove Lemma 11, we need to get a better handle on the set $[0, 1]$. To do so, we exploit more directly the geometry of the braid group.

Definition 12. A positive braid A is called a **positive permutation braid** if it can be drawn as a geometric braid in which every pair of strings crosses at most once. In this case, we say $A \in S_n^+$.

The choice of terminology is motivated by the following Proposition, which establishes a bijective correspondence between S_n^+ and S_n.

Proposition 13. *If $A_1, A_2 \in S_n^+$ induce the same permutation on their strings, then $A_1 = A_2$. For any $\pi \in S_n$, there exists a braid $A_\pi \in S_n^+$ which induces the permutation π on its strings.*

Proof (Sketch). A full proof is available in [EM]. Suppose that $A_1, A_2 \in S_n^+$ induce the same permutation on their strings. By definition, two strings i, j have at most one crossing point in which string j passes in front of string i if $i < j$. Hence, we can realize the braid so that each string lies in a vertical plane, with the left most (at the top of the braid) string at the furthest back level and each other string at successively higher levels. Because A_1 and A_2 induce the same permutation on their strings, we can isotope A_1 into A_2 keeping each string in its plane, which shows that $A_1 = A_2$ as braids.

Now, suppose we have a permutation π. We need to find a braid A_π that induces the permutation π in which each pair of strings crosses at most once. Select n points on the top and bottom of a rectangle and connected them with n lines joining the i-th point at the top with the $\pi(i)$-th point at the bottom such that two lines cross at most once. It is clear we can always do this for a permutation. Now, convert this diagram into a braid by turning each crossing into a positive braid crossing. See Figure 3.3. □

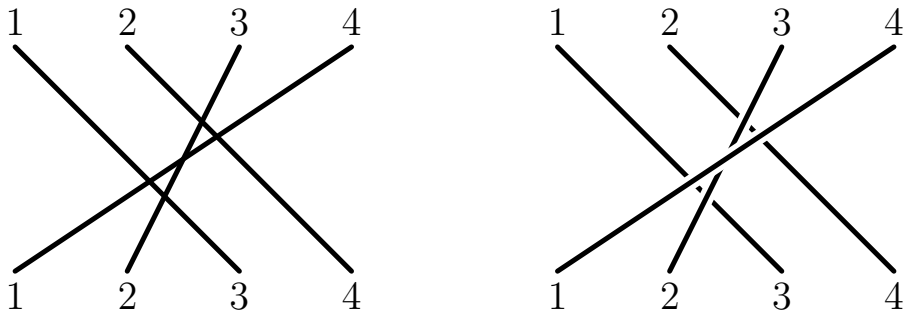


Figure 3.3: Diagram associated to (1324) becomes a braid in S_n^+

The following provides a usefully criterion for computing the starting set of a permutation braid.

Lemma 14. *For $A_\pi \in S_n^+$, the following are equivalent:*

1. $i \in S(A_\pi)$,
2. *Strings i and $i+1$ cross in A_π,*
3. $\pi(i+1) < \pi(i)$.

Proof. (2) and (3) are equivalent since strings cross at most once. If we have $A_\pi = \sigma_i A'$ for a positive A' (that is to say, if $i \in S(A_\pi)$), then (2) follows since the strings $i, i+1$ cross in σ_i. Conversely, if strings i and $i+1$ cross in A_π, then we can draw a diagram of this permutation in which this crossing happens first. Converting this diagram into a braid yields the other implication. □

Corollary 15. *Let $A \in S_n^+$. Then $\sigma_i A \in S_n^+ \iff i \notin S(A)$.*

Proof. The strings i and $i+1$ of $\sigma_i A$ cross once if $i \notin S(A)$ and twice otherwise. All other pairs cross at most once. The result then follows directly from Lemma 14 □

A similar result holds for the finishing set by applying Corollary 15 to $\mathrm{rev}\, A$. As a result of our geometric detour, we can now establish the following

Proposition 16. *The subsets $[0,1]$ and S_n^+ of B_n are the same.*

Proof. Recall that $[0,1]$ simply means the set of all braids B such that $e \leq B \leq \Delta$. Clearly, $\Delta \in S_n^+$ since every pair of strings crosses exactly once. Further, if $P = AB \in S_n^+$ where $A, B \geq e$, then any pair of strings in A crosses at most once in P and hence crosses at most once in A. So $A \in S_n^+$. Similarly, $B \in S_n^+$. Now, by Lemma 5, for every braid $A \in [0,1]$, there exists a positive braid B such that $AB = \Delta$, hence $A \in S_n^+$.

Conversely, say that $A = A_\pi \in S_n^+$. Let $\delta \in S_n$ be the permutation associated to Δ and let $\rho = \pi^{-1}\delta$. Then $A_\pi A_\rho$ is a positive braid with permutation $\pi\rho = \delta$. It is enough to show that $A_\pi A_\rho \in S_n^+$, for then, by Proposition 13, $A_\pi A_\rho = A_\delta = \Delta$. It follows that since $e \leq A_\pi A_\rho = \Delta$, $A_\pi \leq \Delta$, so $A_\pi \in [0, 1]$. Now, any pair of strings in $A_\pi A_\rho$ can cross at most twice since $A_\pi, A_\rho \in S_n^+$. But, the permutation associated to $A_\pi A_\rho$ is δ so each pair of strings crosses an odd number of times and so crosses exactly once. To see that any braid with permutation δ need have each strand cross an odd number of times, notice that $\delta(i) = n - i$, which induces one crossing of each pair of strands. All other crossings must come in pairs, so the total number of crossings is odd. It follows that $A_\pi A_\rho \in S_n^+$. □

The proof of the Main Lemma relies on the following technical result which we shall not prove. It was first proved by Garside and is referenced in [EM].

Lemma 17 (Garside's Lemma)**.** *Let $P = \sigma_i P_1 = \sigma_j P_2$ where P_1, P_2 are positive braids. Then there exists a positive braid P_3 such that $P = (\sigma_i * \sigma_j)P_3$, where*

$$\sigma_i * \sigma_j = \begin{cases} \sigma_i & i = j \\ \sigma_i\sigma_j & |i - j| \geq 2 \\ \sigma_i\sigma_j\sigma_i & |i - j| = 1 \end{cases}$$

With this machinery in place, we can finally turn to the proof of Lemma 11.

Proof of Lemma 11 (Main Lemma). We first show the existence of a left-weighted factorization $P = A_1P_1$ with $A_1 \in S_n^+$. Consider all positive factorizations $P = AB$, with $A \in S_n^+$. Such factorizations surely exist as each generator $\sigma_i \in S_n^+$. Choose one in which $\operatorname{wt} A$ is maximal. If $S(B) \not\subset F(A)$, then choose $i \in S(B) - F(A)$. By Corollary 15, $A' = A\sigma_i \in S_n^+$. We can write $B = \sigma_i B'$, with B' positive. This yields a new positive factorization $P = A'B'$ with $A' \in S_n^+$ and $\operatorname{wt} A' > \operatorname{wt} A$, which is a contradiction. Hence, the selected factorization must be left-weighted. Write this factorization as $P = A_1P_1$.

Now we turn to the universality. In partiuclar, we will show that given any other positive factorization $P = AB$ with $A \in S_n^+$, we have $A_1 = AQ$ for some positive braid Q. Suppose not. Then there exists factorizations $P = C\sigma_i B'$ with $C\sigma_i \in S_n^+$ and such that $C \in S_n+$ is a subfactor of A_1 but $C\sigma_i$ is not. Choose such a factorization with maximal weight $\operatorname{wt} C$ and write $A_1 = CQ$. Now $C\sigma_i B'$ is a positive factorization of P, so by the maximality of $\operatorname{wt} A_1$, we have $\operatorname{wt} A_1 \geq \operatorname{wt} C\sigma_i > \operatorname{wt} C$. Hence $Q \neq e$ so we can choose $j \in S(Q)$. Then $C\sigma_j \leq A_1$, so $C\sigma_j \in S_n^+$ (since $A_1 \in S_n^+$ and $S_n^+ = [0, 1]$). This gives a factorization as $P = C\sigma_j B''$. By Garside's lemma applied to $B = \sigma_i B' = \sigma_j B''$, we can find a positive braid B''' so that $P = C(\sigma_i * \sigma_j)B'''$.

Since $C\sigma_j \in S_n^+$, $\sigma_j \notin F(C)$ by Corollary 15. Similarly, $\sigma_i \notin F(C)$. An extension of the proof of Corollary 15 to the case of $\sigma_i * \sigma_j$ shows that $C(\sigma_i * \sigma_j) \in S_n^+$. We omit the technical details. As a consequence, we have a factorization of P with a larger subfactor (at least including $C\sigma_j$) in common with A_1, while $C(\sigma_i * \sigma_j)$ itself is not a subfactor of A_1 (by choice of σ_i). This contradicts the maximality of $\operatorname{wt} C$.

It follows that the left-weighted factorization $P = A_1P_1$ is unique. For, if $P = AB$ is another left-weighted factorization, then we have $A_1 = AQ$ for a positive braid Q. If $Q \neq e$, choose $i \in S(Q)$. By Corollary 15, $i \notin F(A)$ since $A\sigma_i \leq A_1 \in S_n^+$. However, $B = QP_1$, so $i \in S(B)$ and the resulting factorization is not left-weighted, which yields a contradiction. Hence $Q = e$ and $A = A_1$. □

Corollary 18. *Let P be a positive braid with left-weighted factorization $P = A_1P_1$ and $A_1 \in S_n^+$. Then $S(A_1) = S(P)$.*

Proof. Clearly $S(A_1) \subset S(P)$. On the other hand, let $i \in S(P)$. Then $P = \sigma_i B$ for some positive braid P. By the main lemma, we have $A_1 = \sigma_i Q$ for some positive braid Q. Hence $i \in S(A_1)$. □

The main theorem is an easy corollary of the main lemma. Recall that we seek to prove the following

Theorem 19 (Main Theorem)**.** *Let P be a positive braid. Then there is a unique expression $P = A_1A_2 \dots A_k$ with $A_i \in [0,1]$, $A_k \neq e$, and $S(A_{i+1}) \subset F(A_i)$ for each i.*

Proof. Let $P = P_0$ and let $P_0 = A_1P_1$ be the unique left-weighted factorization with $A_1 \in [0,1]$. By induction, factorize $P_i = A_{i+1}P_{i+1}$ with $S(P_{i+1}) \subset F(A_i)$. By Corollary 18, $S(P_{i+1}) = S(A_{i+1})$, so $S(A_{i+1}) \subset F(A_i)$. It is clear this this process terminates since the weight of P_i is strictly less than the weight of P_{i-1}, while the total weight of P is finite. □

3.4 An Algorithm for Braid Normal Form

Suppose that we have a braid word B. We seek to write $B = \Delta^r P'$ where P' is a positive braid, but $P' \not\geq \Delta$ and to compute the left-weighted canonical form of P' as a sequence of permutations. Again, we closely follow the discussion of [EM].

In order to start the algorithm, we need to write the braid in the form $\Delta^r P'$. We can do this by rewriting each σ_i^{-1} as $\Delta^{-1}\sigma_i^*$, where $\sigma_i^* \in [0,1]$ and then collect factors of Δ to the left by repeatedly using $A\Delta = \Delta\tau(A)$. We also need to write P' as a sequence of permutation braids, which is easy since each generator σ_i is itself a permutation braid.

Now, suppose that $B = \Delta^r P'$ and P' is the product $B_1 \dots B_k$ of positive permutation braids. Find the sets $F(B_i)$ and $S(B_i)$. If $S(B_{i+1}) \subset F(B_i)$ for each i, then we have found the canonical form by the uniqueness of the Main Theorem (except for some final factors of e, which can be ignored). Incorporate any initial factors of Δ into the power r and return r and the sequence of permutations defining the permutation braids. If not, then find the first i such that $S(B_{i+1}) \not\subset F(B_i)$ and choose $j \in S(B_{i+1})$ so that $j \notin F(B_i)$. Consider the braids $C_i = B_i\sigma_j$ and $C_{i+1} = \sigma_j^{-1}B_{i+1}$. $C_i \in S_n^+$ by Corollary 15. $C_{i+1} = \sigma_j^{-1}B_{i+1} \in S_n^+$ since $e \leq \sigma_j^{-1}B_{i+1} \leq B_{i+1}$ ($j \in S(B_{i+1})$). Replace B_i, B_{i+1} with C_i, C_{i+1} and iterate the process. This completes the algorithm.

The algorithm terminates because at each stage, the weight of the permutation subwords increases and there are a finite number of words of a given total weight.

3.5 Applications: Braids and Diffie-Hellman Key Exchange

Given a braid normal form, we can implement the braid group on a computer by representing braids as normalized words. We can implement the group operation simply as concatenation of words followed by normalization. This enables us to easily determine braid equivalence and hence implement cryptographic algorithms based on braids. One example is a key exchange scheme that is very similar to the familiar Diffie-Hellman scheme.

The goal is to securely construct a shared piece of information known only by two parties (traditionally named **A**lice and **B**ob). This piece of information can then be used, in conjunction with other cryptographic systems, to securely transmit messages between the two parties.

Let LB_n denote the subgroup of B_n generated by $\sigma_1, \dots, \sigma_m$, where $m = \lfloor n/2 \rfloor$. Similarly, let UB_n denote the subgroup generated by $\sigma_{m+1}, \dots, \sigma_{n-1}$. It is clear (either geometrically or by the far commutativity relation) that each braid in LB_n commutes with every braid in UB_n. Suppose that a public braid p is known by both parties (and potential adversaries). Further suppose that **A** has a secret braid $r \in LB_n$ and **B** has a secret braid $s \in UB_n$. Carry out the following sequence of exchanges:

1. **A** computes rpr^{-1} and sends it to **B**
2. **B** computes sps^{-1} and sends it to **A**
3. **A** computes $\tau_A = r(sps^{-1})r^{-1}$
4. **B** computes $\tau_B = s(rpr^{-1})s^{-1}$

Since r and s commute, we have a shared secret $\tau = \tau_A = \tau_B$. In order for an adversary to determine τ, they need to solve a variant of the conjugator search problem, which is believed to be

difficult given a suitable choice of braid p and group size n. The conjugator search problem asks, given braids A, B, to find a braid Q such that $A = QBQ^{-1}$. The problem here is a variant on this problem since the adversary knows two braids conjugate to p, namely rpr^{-1} and sps^{-1} and also knows that $r \in LB_n$ (resp. $s \in UB_n$).

Initial research suggested that this problem should be very difficult, however subsequent efforts have shown the Braid Group has more internal structure than initially suspected and that, hence, it may be possible to some the conjugator search problem. More details about the difficulty of these problems and various other cryptographic systems employing braids can be found in [De].

3.6 Conclusion

Although research interest in the braid group for cryptographic purposes has slowed in recent years, the group may hold cryptographic promise. The traditional assumption is that a randomly chosen braid in a group on the order of B_{80} is ideal for a public-key. Recent research has shown this not to be the case [De]. Still, it may be possible to construct a secure system by choosing suitable braids p, r, s. More research is necessary to determine whether such a system can be made secure.

Even if the braid group turns out not to yield a secure cryptography system, the group and its word problem are interesting from a strictly mathematical perspective. The unique blend of topology and algebra employed in analyzing braids gives the subject a refreshing and beautiful flavor.

References

[Bi] Joan Birman: *Braids, Links, and Mapping Class Groups*. Princeton: Princeton University Press 1975.

[De] Patrick Dehorney: Braid-based cryptography, *Cont. Math.*, **360** (2000), 5–33.

[El] James H. Ellis: The Possibility of Secure Non-Secret Digital Encryption, CESG Report, (1970).

[EM] Elsayed Elrifai and Hugh Morton: Algorithms for positive braids, *Q. J. Math.*, **45** (1994), 479.

[Ha] Vagn Lundsgaard Hansen: *Braids and Coverings: Selected Topics*. Cambridge: Cambridge University Press 1989.

[Ma] Vassily Manturov: *Knot Theory*. Boca Raton: CRC Press 2004.

[NC] Michael A. Nielsen and Isaac L. Chuang: *Quantum Computation and Quantum Information*. Cambridge: Cambridge University Press 2000.

STUDENT ARTICLE

4

Young Tableaux and the Representations of the Symmetric Group

Yufei Zhao[†]
Massachusetts Institute of Technology '10
Cambridge, MA 02139
yufeiz@mit.edu

Abstract

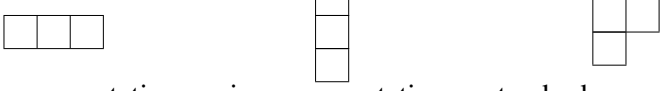
We explore an intimate connection between Young tableaux and representations of the symmetric group. We describe the construction of Specht modules which are irreducible representations of S_n, and also highlight some interesting results such as the branching rule and Young's rule. Some knowledge of basic representation theory is assumed.

4.1 Introduction

In this article, we explore a connection between representations of the symmetric group S_n and combinatorial objects called Young tableaux. We define Young tableaux in Section 4.2, but for now, it suffices to say that they are fillings of a certain configuration of boxes with entries from $\{1, 2, \ldots, n\}$, an example of which is shown below.

1	2	4
3	5	6
7	8	
9		

So how are representations of S_n related to Young tableau? It turns out that there is a very elegant description of irreducible representations of S_n through Young tableaux. Let us have a glimpse of the results. Recall that there are three irreducible representations of S_3. It turns out that they can be described using the set of Young diagrams with three boxes. The correspondence is illustrated below.



trivial representation sign representation standard representation

It is true in general that the irreducible representations of S_n can be described using Young diagrams of n boxes! Furthermore, we can describe a basis of each irreducible representation using standard Young tableaux, which are numberings of the boxes of a Young diagram with $1, 2, \ldots, n$ such that the rows and columns are all increasing. For instance, the bases of the standard representation of S_3 correspond to the following two standard Young tableaux:

1	2
3	

1	3
2	

[†]Yufei Zhao, Massachusetts Institute of Technology '10, is a mathematics and computer science major. His favorite mathematical area is combinatorics, but he also enjoys algebra and number theory. He is regularly involved with the training of the Canadian team for the International Math Olympiad.

The dimension of the irreducible representations can be easily computed from its Young diagram through a result known as the hook-length formula, as we explain in Section 4.4.

There are many other surprising connections between Young tableaux and representations of S_n, one of which is the following. Suppose we have an irreducible representation in S_n and we want to find its induced representation in S_{n+1}. It turns out that the induced representation is simply the direct sum of all the representations corresponding to the Young diagrams obtained by adding a new square to the original Young diagram! For instance, the induced representation of the standard representation from S_3 to S_4 is simply

$$\mathrm{Ind}_{S_3}^{S_4} \begin{array}{|c|c|}\hline & \\ \hline \\ \cline{1-1} \end{array} = \begin{array}{|c|c|c|}\hline & & \\ \hline \\ \cline{1-1} \end{array} \oplus \begin{array}{|c|c|}\hline & \\ \hline & \\ \hline \end{array} \oplus \begin{array}{|c|c|}\hline & \\ \hline \\ \cline{1-1} \end{array}.$$

Similarly, the restricted representation can be found by removing a square from the Young diagram:

$$\mathrm{Res}_{S_2} \begin{array}{|c|c|}\hline & \\ \hline \\ \cline{1-1} \end{array} = \begin{array}{|c|c|}\hline & \\ \hline \end{array} \oplus \begin{array}{|c|}\hline \\ \hline \\ \cline{1-1} \end{array}.$$

In this paper, we describe the connection between Young tableaux and representations of S_n. The goal is to attract readers to the subject by showing a selection of very elegant and surprising results. Most proofs are omitted, but those who are interested may find them in [Fu], [FH], or [Sa]. We assume familiarity with the basics of group representations, including irreducible representations and characters. Induced representations are used in Section 4.5. For references on group representations, see [FH], [Sa], or [Se].

In Section 4.2, we introduce Young diagrams and Young tableaux. In Section 4.3, we introduce tabloids and use them to construct a representation of S_n known as the permutation module M^λ. However, permutation modules are generally reducible. In Section 4.4, we construct irreducible representations of S_n known as Specht modules S^λ. Specht modules S^λ correspond bijectively to Young diagrams λ and they form a complete list of irreducible representations. In Section 4.5, we discuss the Young lattice and the branching rule, which are used to determine the induced and restricted representations of S^λ. Finally, in Section 4.6, we introduce Kostka numbers and state a result concerning the decomposition of permutation modules into the irreducible Specht modules.

4.2 Young Tableaux

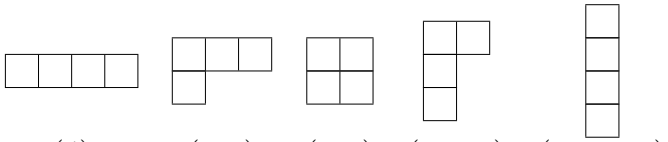
First we need to settle some definitions and notations regarding partitions and Young diagrams.

Definition 1. A **partition** of a positive integer n is a sequence of positive integers $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_l)$ satisfying $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_l > 0$ and $n = \lambda_1 + \lambda_2 + \cdots + \lambda_l$. We write $\lambda \vdash n$ to denote that λ is a partition of n.

For instance, the number 4 has five partitions: $(4), (3,1), (2,2), (2,1,1), (1,1,1,1)$. We can also represent partitions pictorially using Young diagrams as follows.

Definition 2. A **Young diagram** is a finite collection of boxes arranged in left-justified rows, with the row sizes weakly decreasing.[1] The Young diagram associated to the partition $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_l)$ is the one that has l rows, and λ_i boxes on the ith row.

For instance, the Young diagrams corresponding to the partitions of 4 are



(4) $(3,1)$ $(2,2)$ $(2,1,1)$ $(1,1,1,1)$

Since there is a clear one-to-one correspondence between partitions and Young diagrams, we use the two terms interchangeably, and we will use Greek letters λ and μ to denote them.

A Young tableau is obtained by filling the boxes of a Young diagram with numbers.

[1] The notation used here is known as the *English notation*. Most Francophones, however, use the *French notation*, which is the upside-down form of the English notation. E.g. $(3,1)$ as $\begin{array}{|c|}\hline \\ \hline \end{array}\begin{array}{|c|c|c|}\hline & & \\ \hline \end{array}$.

Definition 3. Suppose $\lambda \vdash n$. A **(Young) tableau t of shape λ**, is obtained by filling in the boxes of a Young diagram of λ with $1, 2, \ldots, n$, with each number occurring exactly once. In this case, we say that t is a λ-tableau.

For instance, here are all the tableaux corresponding to the partition $(2, 1)$:

$$\begin{array}{|c|c|}\hline 1&2\\\hline 3\\\cline{1-1}\end{array} \quad \begin{array}{|c|c|}\hline 2&1\\\hline 3\\\cline{1-1}\end{array} \quad \begin{array}{|c|c|}\hline 1&3\\\hline 2\\\cline{1-1}\end{array} \quad \begin{array}{|c|c|}\hline 3&1\\\hline 2\\\cline{1-1}\end{array} \quad \begin{array}{|c|c|}\hline 2&3\\\hline 1\\\cline{1-1}\end{array} \quad \begin{array}{|c|c|}\hline 3&2\\\hline 1\\\cline{1-1}\end{array}$$

Definition 4. A **standard (Young) tableau** is a Young tableaux whose the entries are increasing across each row and each column.

The only standard tableaux for $(2, 1)$ are

$$\begin{array}{|c|c|}\hline 1&2\\\hline 3\\\cline{1-1}\end{array} \quad \text{and} \quad \begin{array}{|c|c|}\hline 1&3\\\hline 2\\\cline{1-1}\end{array}.$$

Here is another example of a standard tableau:

$$\begin{array}{|c|c|c|}\hline 1&2&4\\\hline 3&5&6\\\hline 7&8\\\cline{1-2} 9\\\cline{1-1}\end{array}.$$

The definitions that we use here are taken from [Sa], however, other authors have different conventions. For instance, in [Fu], a Young tableau is a filling which is weakly increasing across each row and strictly increasing down each column, but may have repeated entries. We call such tableaux **semistandard** and we use them in Section 4.6.

Before we move on, let us recall some basic facts about permutations. Every permutation $\pi \in S_n$ has a decomposition into disjoint cycles. For instance $(123)(45)$ denotes the permutation that sends $1 \to 2 \to 3 \to 1$ and swaps 4 and 5 (if $n > 5$, then by convention the other elements are fixed by π). The **cycle type** of π is the partition whose parts are the lengths of the cycles in the decomposition. So $(123)(45) \in S_5$ has cycle type $(3, 2)$. It is a basic result that two elements of S_n are conjugates if and only if they have the same cycle type. The easiest way to see this is to consider conjugation as simply a relabeling of the elements when the permutation is written in cycle notation. Indeed, if

$$\pi = (a_1 a_2 \ldots a_k)(b_1 b_2 \ldots b_l) \cdots,$$

and σ sends x to x', then

$$\sigma\pi\sigma^{-1} = (a_1' a_2' \ldots a_k')(b_1' b_2' \ldots b_l') \cdots.$$

This means that the conjugacy classes of S_n are characterized by the cycle types, and thus they correspond to partitions of n, which are equivalent to Young diagrams of size n. Recall from representation theory that the number of irreducible representations of a finite group is equal to the number of its conjugacy classes. So our goal for the next two sections is to construct an irreducible representation of S_n corresponding to each Young diagram.

4.3 Tabloids and the Permutation Module M^λ

We would like to consider certain permutation representations of S_n. There is the obvious one: the permutation action of S_n on the elements $\{1, 2, \ldots, n\}$, which extends to the **defining representation**. In this section, we construct other representations of S_n using equivalence classes of tableaux, known as tabloids.

Definition 5. Two λ-tableaux t_1 and t_2 are **row-equivalent**, denoted $t_1 \sim t_2$, if the corresponding rows of the two tableaux contain the same elements. A **tabloid** of shape λ, or λ-tabloid is such an equivalence class, denoted by $\{t\} = \{t_1 \mid t_1 \sim t\}$ where t is a λ-tabloid. The tabloid $\{t\}$ is drawn as the tableaux t without vertical bars separating the entries within each row.

For instance, if

$$t = \begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 \\ \cline{1-1}\end{array}$$

then $\{t\}$ is the tabloid drawn as

$$\begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 \\ \cline{1-1}\end{array}$$

which represents the equivalence class containing the following two tableaux:

$$\begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 \\ \cline{1-1}\end{array} \quad \begin{array}{|c|c|}\hline 2 & 1 \\ \hline 3 \\ \cline{1-1}\end{array}$$

The notation is suggestive as it emphasizes that the order of the entries within each row is irrelevant, so that each row may be shuffled arbitrarily. For instance:

$$\begin{array}{|c|c|c|}\hline 1 & 4 & 7 \\ \hline 3 & 6 \\ \cline{1-2} 2 & 5 \\ \cline{1-2}\end{array} = \begin{array}{|c|c|c|}\hline 4 & 7 & 1 \\ \hline 6 & 3 \\ \cline{1-2} 2 & 5 \\ \cline{1-2}\end{array} \neq \begin{array}{|c|c|c|}\hline 4 & 7 & 1 \\ \hline 6 & 5 \\ \cline{1-2} 2 & 3 \\ \cline{1-2}\end{array} \neq \begin{array}{|c|c|c|}\hline 4 & 7 & 1 \\ \hline 2 & 3 \\ \cline{1-2} 6 & 5 \\ \cline{1-2}\end{array}$$

We want to define a representation of S_n on a vector space whose basis is exactly the set of tabloids of a given shape. We need to find a way for elements of S_n to act on the tabloids. We can do this in the most obvious manner, that is, by letting the permutations permutate the entries of the tabloid. For instance, the cycle $(1\ 2\ 3) \in S_3$ acts on a tabloid by changing replacing its "1" by a "2", its "2" by a "3", and its "3" by a "1", as shown below:

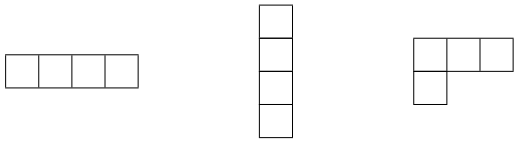
$$(1\ 2\ 3)\begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 \\ \cline{1-1}\end{array} = \begin{array}{|c|c|}\hline 2 & 3 \\ \hline 1 \\ \cline{1-1}\end{array}$$

We should check that this action is well defined, that is, if t_1 and t_2 are row-equivalent, so that $\{t_1\} = \{t_2\}$, then the result of permutation should be the same, that is, $\pi\{t_1\} = \pi\{t_2\}$. This is clear, as π simply gives the instruction of moving some number from one row to another.

Now that we have defined a way for S_n to act on tabloids, we are ready to define a representation of S_n. Recall that a representation of a group G on a complex vector space V is equivalent to extending V to a $\mathbb{C}[G]$-module, so we often use the term **module** to describe representations.

Definition 6. Suppose $\lambda \vdash n$. Let M^λ denote the vector space whose basis is the set of λ-tabloids. Then M^λ is a representation of S_n known as the **permutation module corresponding to $\boldsymbol{\lambda}$**.

Let us show a few example of permutation modules. We see that the M^λ corresponding to the following Young diagrams are in fact familiar representations.



Example 7. Consider $\lambda = (n)$. We see that M^λ is the vector space generated by the single tabloid

$$\begin{array}{|c|c|c|c|}\hline 1 & 2 & \cdots & n \\ \hline\end{array}.$$

Since this tabloid is fixed by S_n, we see that $M^{(n)}$ is the one-dimensional trivial representation.

Example 8. Consider $\lambda = (1^n) = (1, 1, \ldots, 1)$. Then a λ-tabloid is simply a permutation of $\{1, 2, \ldots, n\}$ into n rows and S_n acts on the tabloids by acting on the corresponding permutation. It follows that $M^{(1^n)}$ is isomorphic to the regular representation $\mathbb{C}[S_n]$.

Example 9. Consider $\lambda = (n-1, 1)$. Let $\{t_i\}$ be the λ-tabloid with i on the second row. Then M^λ has basis $\{t_1\}, \{t_2\}, \ldots, \{t_n\}$. Also, note that the action of $\pi \in S_n$ sends t_i to $t_{\pi(i)}$. And so $M^{(n-1,1)}$ is isomorphic to the defining representation $\mathbb{C}\{1, 2, \ldots, n\}$. For example, in the $n = 4$ case, the representation $M^{(3,1)}$ has the following basis:

$$t_1 = \begin{array}{|c|c|c|c|}\hline 2 & 3 & 4 \\ \hline 1 \\ \cline{1-1}\end{array}, \quad t_2 = \begin{array}{|c|c|c|c|}\hline 1 & 3 & 4 \\ \hline 2 \\ \cline{1-1}\end{array}, \quad t_3 = \begin{array}{|c|c|c|c|}\hline 1 & 2 & 4 \\ \hline 3 \\ \cline{1-1}\end{array}, \quad t_4 = \begin{array}{|c|c|c|c|}\hline 1 & 2 & 3 \\ \hline 4 \\ \cline{1-1}\end{array}.$$

Now we consider the dimension and characters of the representation M^λ. First, we shall give a formula for the number of tabloids of each shape.

Proposition 10. *If $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_l)$,*

$$\dim M^\lambda = \frac{n!}{\lambda_1!\lambda_2! \cdots \lambda_l!}.$$

We leave the proof of this proposition to the readers. It is a simple combinatorial exercise of counting the number of λ-tabloids.

Now we give a formula for the characters of M^λ.

Proposition 11. *Suppose $\lambda = (\lambda_1, \ldots, \lambda_l), \mu = (\mu_1, \ldots, \mu_m)$ are partitions of n. The character of M^λ evaluated at an element of S_n with cycle type μ is equal to the coefficient of $x_1^{\lambda_1} x_2^{\lambda_2} \cdots x_l^{\lambda_l}$ in*

$$\prod_{i=1}^{m} (x_1^{\mu_i} + x_2^{\mu_i} + \cdots + x_l^{\mu_i}).$$

To prove this formula, note that since M^λ can be realized as a permutation representation on the λ-tabloids, its character at an element $\pi \in S_n$ is equal to the number of tabloids fixed by π. The rest of the proof consists of a simple generating function argument, which we leave to the readers.

Note that Proposition 10 also follows as a corollary to the above result. Indeed, the dimension of a representation is simply the value of the character at the identity element, which has cycle type $\mu = (1^n)$. So Proposition 11 tells us that the dimension of M^λ is the coefficient of $x_1^{\lambda_1} x_2^{\lambda_2} \cdots x_l^{\lambda_l}$ in $(x_1 + \cdots + x_n)^n$, which is equal to $\dim M^\lambda = \frac{n!}{\lambda_1!\lambda_2! \cdots \lambda_l!}$ by the multinomial expansion formula.

Example 12. Let us compute the full list of the characters of the permutation modules for S_4. The character at the identity element is equal to the dimension, and it can found through Proposition 10. For instance, the character of $M^{(2,1,1)}$ at $e \in s_4$ is $4!/2! = 12$.

Say we want to compute the character of $M^{(2,2)}$ at the permutation (12), which has cycle type $(2, 1, 1)$. Using Proposition 11, we see that the character is equal to the coefficient of $x_1^2 x_2^2$ in $(x_1^2 + x_2^2)(x_1 + x_2)^2$, which is 2. Other characters can be similarly computed, and the result is shown in the following table.

permutation cycle type	e $(1,1,1,1)$	(12) $(2,1,1)$	$(12)(34)$ $(2,2)$	(123) $(3,1)$	(1234) (4)
$M^{(4)}$	1	1	1	1	1
$M^{(3,1)}$	4	2	0	1	0
$M^{(2,2)}$	6	2	2	0	0
$M^{(2,1,1)}$	12	2	0	0	0
$M^{(1,1,1,1)}$	24	0	0	0	0

Note that in the above example, we did *not* construct the character table for S_4, as all the M^λ are in fact reducible with the exception of $M^{(4)}$. In the next section, we take a step further and construct the irreducible representations of S_n.

4.4 Specht Modules

In the previous section, we constructed representations M^λ of S_n known as permutation modules. In this section, we consider an irreducible subrepresentation of M^λ that corresponds uniquely to λ.

The group S_n acts on the set of Young tableaux in the obvious manner: for a tableaux t of size n and a permutation $\sigma \in S_n$, the tableaux σt is the tableaux that puts the number $\pi(i)$ to the box where t puts i. For instance,

$$(1\,2\,3)(4\,5)\begin{array}{|c|c|c|c|}\hline 1&2&4&5\\\hline 3&6\\\cline{1-2}7\\\cline{1-1}\end{array} = \begin{array}{|c|c|c|c|}\hline 2&3&5&4\\\hline 1&6\\\cline{1-2}7\\\cline{1-1}\end{array}\,.$$

Observe that a tabloid is fixed by the permutations which only permute the entries of the rows among themselves. These permutations form a subgroup of S_n, which we call the row group. We can similarly define the column group.

Definition 13. For a tableau t of size n, the **row group** of t, denoted R_t, is the subgroup of S_n consisting of permutations which only permutes the elements within each row of t. Similarly, the **column group** C_t is the the subgroup of S_n consisting of permutations which only permutes the elements within each column of t.

For instance, if

$$t = \begin{array}{|c|c|c|}\hline 4&1&2\\\hline 3&5\\\cline{1-2}\end{array}$$

then

$$R_t = S_{\{1,2,4\}} \times S_{\{3,5\}}, \quad \text{and} \quad C_t = S_{\{3,4\}} \times S_{\{1,5\}} \times S_{\{2\}}.$$

Let us select certain elements from the space M^λ that we use to to span a subspace.

Definition 14. If t is a tableau, then the associated **polytabloid** is

$$e_t = \sum_{\pi \in C_t} \operatorname{sgn}(\pi)\pi\{t\}.$$

So we can find e_t by summing all the tabloids that come from column-permutations of t, taking into account the sign of the column-permutation used. For instance, if

$$t = \begin{array}{|c|c|c|}\hline 4&1&2\\\hline 3&5\\\cline{1-2}\end{array},$$

then

$$e_t = \begin{array}{|c|c|c|}\hline 4&1&2\\\hline 3&5\\\cline{1-2}\end{array} - \begin{array}{|c|c|c|}\hline 3&1&2\\\hline 4&5\\\cline{1-2}\end{array} - \begin{array}{|c|c|c|}\hline 4&5&2\\\hline 3&1\\\cline{1-2}\end{array} + \begin{array}{|c|c|c|}\hline 3&5&2\\\hline 4&1\\\cline{1-2}\end{array}\,.$$

Now, through the following technical lemma, we see that S_n acts on the set of polytabloids.

Lemma 15. *Let t be a tableau and π be a permutation. Then $e_{\pi t} = \pi e_t$.*

Proof. First observe that $C_{\pi t} = \pi C_t \pi^{-1}$, which can be viewed as a "relabeling" similar to the discussion at the end of Section 4.2. Then, we have

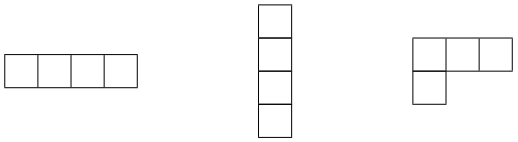
$$\begin{aligned} e_{\pi t} &= \sum_{\sigma \in C_{\pi t}} \operatorname{sgn}(\sigma)\sigma\{\pi t\} = \sum_{\sigma \in \pi C_t \pi^{-1}} \operatorname{sgn}(\sigma)\sigma\{\pi t\} \\ &= \sum_{\sigma' \in C_t} \operatorname{sgn}(\pi\sigma'\pi^{-1})\pi\sigma'\pi^{-1}\{\pi t\} = \pi \sum_{\sigma' \in C_t} \operatorname{sgn}(\sigma')\sigma'\{t\} = \pi e_t. \end{aligned}$$

□

Now we are ready to extract an irreducible subrepresentation from M^λ.

Definition 16. For any partition λ, the corresponding **Specht module**, denoted S^λ, is the submodule of M^λ spanned by the polytabloids e_t, where t is taken over all tableaux of shape λ.

Again, let us look at a few examples. We see that the Specht modules corresponding to the following Young diagrams are familiar irreducible representations.



Example 17. Consider $\lambda = (n)$. Then there is only one polytabloid, namely

1	2	$\cdots$	n

Since this polytabloid is fixed by S_n, we see that $S^{(n)}$ is the one-dimensional trivial representation.

Example 18. Consider $\lambda = (1^n) = (1, 1, \ldots, 1)$. Let

$$t = \begin{array}{|c|}\hline 1 \\ \hline 2 \\ \hline \vdots \\ \hline n \\ \hline \end{array}$$

Observe that e_t is a sum of all the λ-tabloids multiplied by the sign of permutation it took to get there. For any other λ-tableau t', we have either $e_t = e_{t'}$ if t' is obtained from t through an even permutation, or $e_t = -e_{t'}$ if t' is obtained from t through an odd permutation. So S^λ is a one-dimensional representation. From Lemma 15 we have $\pi e_t = e_{\pi t} = \mathrm{sgn}(\pi) e_t$. From this we see that $S^{(1^n)}$ is the sign representation.

Example 19. Consider $\lambda = (n-1, 1)$. Continuing the notation from Example 9 where we use $\{t_i\}$ to denote the λ-tabloid with i on the second row, we see that the polytabloids have the form $\{t_i\} - \{t_j\}$. Indeed, the polytabloid constructed from the tableau

i	a	b	$\cdots$
j			

is equal to $\{t_i\} - \{t_j\}$. Let us temporarily use $\mathbf{e}_i$ to denote the tabloid $\{t_i\}$. Then S^λ is spanned by elements of the form $\mathbf{e}_i - \mathbf{e}_j$, and it follows that

$$S^{(n-1,1)} = \{c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \cdots c_n\mathbf{e}_n \mid c_1 + c_2 + \cdots + c_n = 0\}.$$

This is an irreducible representation known as the standard representation. The direct sum of the standard representation and the trivial representation gives the defining representation, that is, $S^{(n-1,1)} \oplus S^{(n)} = M^{(n-1,1)}$.

We know that the S_3 has three irreducible representations: trivial, sign, and standard. These are exactly the ones described above. Furthermore, there are exactly three partitions of 3: (3), $(1,1,1)$, $(2,1)$. So in this case, the irreducible representations are exactly the Specht modules. Amazingly, this is true in general.

Theorem 20. *The Specht modules S^λ for $\lambda \vdash n$ form a complete list of irreducible representations of S_n over $\mathbb{C}$.*

The proof may be found in [Sa]. Recall that at the end of Section 4.2 we noted that the number of irreducible representations of S_n equals the number of Young diagrams with n boxes. This Theorem gives a "natural" bijection between the two sets.

Note that the polytabloids are generally not independent. For instance, as we saw in Example 18, any pair of polytabloids in $S^{(1^n)}$ are in fact linearly dependent. Since we know that S^λ is spanned by the polytabloids, we may ask how to select a basis for vector space from the set of polytabloids. There is an elegant answer to this question: the set of polytabloids constructed from standard tableaux form a basis for S^λ. Recall that a standard tableau is a tableau with increasing rows and increasing columns.

Theorem 21. *Let λ be any partition. The set*

$$\{e_t : t \text{ is a standard } \lambda\text{-tableau}\}$$

forms a basis for S^λ as a vector space.

The proof may be found in Sagan [Sa]. We only sketch an outline here. First, an ordering is imposed on tabloids. If some linear combination of e_t is zero, summed over some standard tableaux t, then by looking at a maximal tabloid in the sum, one can deduce that its coefficient must be zero and conclude that $\{e_t : t \text{ is a standard } \lambda\text{-tableau}\}$ is independent. Next, to prove that the set spans S^λ, a procedure known as the **straightening algorithm** is used to write an arbitrary polytabloid as a linear combination of standard polytabloids.

Now we look at some consequences of the result. Let f^λ denote the number of standard λ-tableaux. Then the following result follows immediately from Theorem 21.

Corollary 22. *Suppose $\lambda \vdash n$, then* $\dim S^\lambda = f^\lambda$.

Let us end this section with a few results concerning f^λ.

Theorem 23. *If n is a positive integer, then*

$$\sum_{\lambda \vdash n} (f^\lambda)^2 = n!$$

where the sum is taken over all partitions of n.

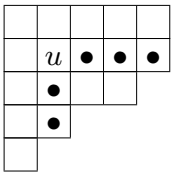
Proof. Recall from representation theory that the sum of the squares of the irreducible representation is equal to the order of the group. This theorem follows from that fact and Corollary 22. □

Theorem 23 also has an elegant combinatorial proof using the celebrated RSK correspondence. See [Fu] or [Sa] for details.

Given the partition λ, the number $\dim S^\lambda = f^\lambda$ can be computed easily using the **hook-length formula** of Frame, Robinson, and Thrall, which we state now.

Definition 24. Let λ be a Young diagram. For a square u in the diagram (denoted by $u \in \lambda$), we define the **hook** of u (or at u) to be the set of all squares directly to the right of u or directly below u, including u itself. The number of squares in the hook is called the **hook-length** of u (or at u), and is denoted by $h_\lambda(u)$.

For example, consider the partition $\lambda = (5, 5, 4, 2, 1)$. The figure on the left shows a typical hook, and the figure on the right shows all the hook-lengths.



9	7	5	4	1
8	6	4	3	1
6	4	2	1	
3	1			
1				

Theorem 25 (Hook-length formula)**.** *Let $\lambda \vdash n$ be a Young diagram. Then*

$$\dim S^\lambda = f^\lambda = \frac{n!}{\prod_{u \in \lambda} h_\lambda(u)}.$$

For instance, from the above example, we get

$$\dim S^{(5,5,4,2,1)} = f^{(5,5,4,2,1)} = \frac{17!}{9 \cdot 8 \cdot 7 \cdot 6^2 \cdot 5 \cdot 4^3 \cdot 3^2 \cdot 2 \cdot 1^5} = 3403400.$$

For proof of the hook-length formula, see [Sa].

Finally, we state a formula for the characters of the representation S^λ.

Theorem 26 (Frobenius formula)**.** *Suppose $\lambda = (\lambda_1, \ldots, \lambda_l), \mu = (\mu_1, \ldots, \mu_m)$ are partitions of n. The character of S^λ evaluated at an element of S_n with cycle type μ is equal to the coefficient of $x_1^{\lambda_1+l-1} x_2^{\lambda_2+l-2} \cdots x_l^{\lambda_l}$ in*

$$\prod_{1 \le i < j \le l} (x_i - x_j) \prod_{i=1}^{m} (x_1^{\mu_i} + x_2^{\mu_i} + \cdots + x_l^{\mu_i}).$$

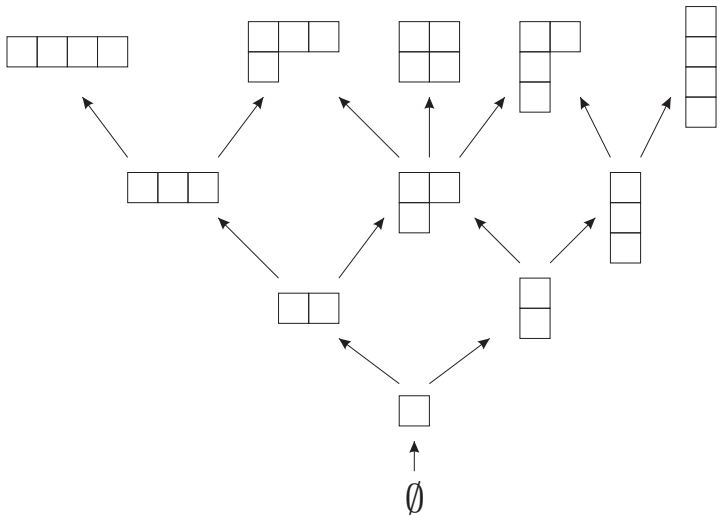
See [FH] for proof. Observe the similarity between the statements of Proposition 10 and the hook-length formula, and also between Proposition 11 and the Frobenius formula. The hook-length formula can also be derived from the Frobenius formula by evaluating the character at the identity element. Again, see [FH] for details.

4.5 Young Lattice and Branching Rule

Now let us consider the relationships between the irreducible representations of S_n and those of S_{n+1}.

Consider the set of all Young diagrams. These diagrams can be partially ordered by inclusion. The resulting partially ordered set is known as **Young's lattice**.

We can represent Young's lattice graphically as follows. Let $\lambda \nearrow \mu$ denote that μ can be obtained by adding a single square to λ. At the nth level, all the Young diagrams with n boxes are drawn. In addition, λ to connected to μ if $\lambda \nearrow \mu$. Here is a figure showing the bottom portion of Young's lattice (of course, it extends infinitely upwards).

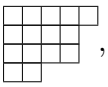


Now we consider the following question: given S^λ a representation of S_n, how can we determine its restricted representation in S_{n-1} and its induced representation in S_{n+1}? There is a beautiful answer to this question, given by Young's branching rule.

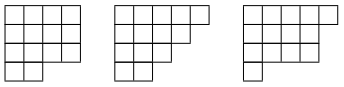
Theorem 27 (Branching Rule)**.** *Suppose $\lambda \vdash n$, then*

$$\mathrm{Res}_{S_{n-1}} S^\lambda \cong \bigoplus_{\mu : \mu \nearrow \lambda} S^\mu \quad \textit{and} \quad \mathrm{Ind}_{S_n}^{S_{n+1}} S^\lambda \cong \bigoplus_{\mu : \lambda \nearrow \mu} S^\mu.$$

For instance, if $\lambda = (5, 4, 4, 2)$, so that

$\lambda =$  ,

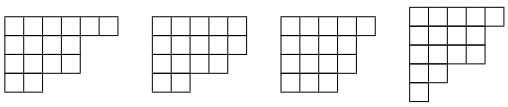
then the diagrams that can be obtained by removing a square are



So

$$\mathrm{Res}_{S_{14}} S^{(5,4,4,2)} = S^{(4,4,4,2)} \oplus S^{(5,4,3,2)} \oplus S^{(5,4,4,1)}.$$

Similarly, the diagrams that can be obtained by adding a square are



So

$$\mathrm{Ind}_{S_{15}}^{S_{16}} S^{(5,4,4,2)} = S^{(6,4,4,2)} \oplus S^{(5,5,4,2)} \oplus S^{(5,4,4,3)} \oplus S^{(5,4,4,2,1)}.$$

The proof of Theorem 27 may be found in [Sa]. We shall only mention that the two parts of the branching rules are equivalent through the Frobenius reciprocity theorem.

There is an interesting way to view this result. If we consider S^λ only as a vector space, then the branching rule implies that

$$S^\lambda \cong \bigoplus_{\mu:\mu\nearrow\lambda} S^\mu \cong \bigoplus_{\nu:\nu\nearrow\mu\nearrow\lambda} S^\nu \cong \cdots \cong \bigoplus_{\emptyset=\lambda^{(0)}\nearrow\lambda^{(1)}\nearrow\cdots\nearrow\lambda^{(n)}=\lambda} S^\emptyset.$$

The final sum is indexed over all upward paths from $\emptyset$ to λ in Young's lattice. Since $S^\emptyset$ is simply an one-dimensional vector space, it follows that we can construct a basis for S^λ where each basis vector corresponds to a upward path in the Young lattice from $\emptyset$ to λ. However, observe that upward paths in the Young lattice from $\emptyset$ to λ correspond to standard λ-tableaux! Indeed, for each standard λ-tableaux, we can associate to it a path in the Young lattice constructed by adding the boxes in order as labeled in the standard tableaux. The reverse construction is similar. As an example, the following path in the Young lattice

corresponds to the following standard tableau

$$\begin{array}{|c|c|c|}\hline 1 & 2 & 4 \\ \hline 3 & 5 \\ \cline{1-2}\end{array}.$$

So we have recovered a basis for S^λ which turned out to be the same as the one found in Theorem 21.

Now, one may object that this argument contains some circular reasoning, namely because the proof of the branching rule (as given in [Sa]) uses Theorem 21, that a basis of S^λ can be found through standard tableaux. This is indeed the case. However, there is an alternative view on the subject, given recently by [VO], in which we start in an abstract algebraic setting with some generalized form of the Young lattice. Then, we can form a basis known as the Gelfand-Tsetlin basis by taking upward paths as we did above. We then specialize to the symmetric group and "discover" the standard tableaux. This means that the standard tableaux in some sense form a "natural" basis for S^λ.

4.6 Decomposition of M^μ and Young's Rule

First, we constructed the permutation modules M^λ, and from it we extracted irreducible subrepresentations S^λ, such that S^λ forms a complete list of irreducible representations of S_n as λ varies over all partitions of n.

Let us revisit M^μ and ask, how does M^μ decompose into irreducible representations. It turns out that M^μ only contains the irreducible S^λ if λ is, in some sense, "greater" than μ. To make this notation more precise, let us define a partial order on partitions of n. (Note that this is not the same as the one used to define Young's lattice!)

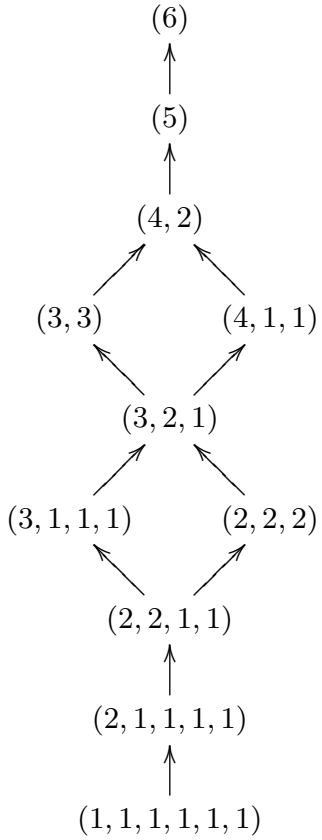
Definition 28. Suppose that $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_l)$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ are partitions of n. Then **λ dominates μ**, written $\lambda \trianglerighteq \mu$, if

$$\lambda_1 + \lambda_2 \cdots + \lambda_i \geq \mu_1 + \mu_2 + \cdots + \mu_i$$

for all $i \geq 1$. If $i > l$ (respectively, $i > m$), then we take λ_i (respectively, μ_i) to be zero.

In other words, $\lambda \trianglerighteq \mu$ if, for every k, the first k rows of the Young diagram of λ contains more squares than that of μ. Intuitively, this means that diagram for λ is short and fat and the diagram for μ is long and skinny.

For example, when $n = 6$, we have $(3, 3) \trianglerighteq (2, 2, 1, 1)$. However, $(3, 3)$ and $(4, 1, 1)$ are incomparable, as neither dominates the other. The dominance relations for partitions of 6 is depicted using the following figure. Such diagrams are known as **Hasse diagrams** and are used to represent partially ordered sets.



Now we can precisely state what we wanted to say at the beginning of the section.

Proposition 29. *M^μ contains S^λ as a subrepresentation if and only if $\lambda \trianglerighteq \mu$. Also, M^μ contains exactly one copy of S^μ.*

We may ask how many copies of S^λ is contained in M^μ. It turns out that this answer has a nice combinatorial interpretation. In order to describe it, we need a few more definitions.

Definition 30. A **semistandard tableau of shape $\boldsymbol{\lambda}$** is an array T obtained by filling in the boxes of λ with positive integers, repetitions allowed, and such that the rows weakly increase and the columns strictly increase. The **content** of T is the composition $\mu = (\mu_1, \mu_2, \ldots, \mu_m)$, where μ_i equals the number of i's in T.

For instance, the semistandard tableau shown below may be seen to have shape $(4, 2, 1)$ and content $(2, 2, 1, 0, 1, 1)$:

$$\begin{array}{|c|c|c|c|}\hline 1&1&2&5\\\hline 2&3\\\cline{1-2} 6\\\cline{1-1}\end{array}\ .$$

The number of semistandard tableau of a given type and content is known as the Kostka number.

Definition 31. Suppose $\lambda, \mu \vdash n$, the **Kostka number $\boldsymbol{K_{\lambda\mu}}$** is the number of semistandard tableaux of shape λ and content μ.

For instance, if $\lambda = (3, 2)$ and $\mu = (2, 2, 1)$, then $K_{\lambda\mu} = 2$ since there are exactly two semistandard tableaux of shape λ and content μ:

$$\begin{array}{|c|c|c|}\hline 1&1&2\\\hline 2&3\\\cline{1-2}\end{array}\qquad\text{and}\qquad\begin{array}{|c|c|c|}\hline 1&1&3\\\hline 2&2\\\cline{1-2}\end{array}\ .$$

We are almost ready to state the result, but let us first make the following observation, whose proof we leave as a combinatorial exercise for the readers.

Proposition 32. *Suppose that $\lambda, \mu \vdash n$. Then $K_{\lambda\mu} \neq 0$ if and only if $\lambda \trianglerighteq \mu$. Also, $K_{\lambda\lambda} = 1$.*

We are now ready to state the result about the decomposition of M^λ into irreducible representations.

Theorem 33 (Young's Rule)**.** $M^\mu \cong \bigoplus_{\lambda \trianglerighteq \mu} K_{\lambda\mu} S^\lambda$.

For instance, from the table above, we see that

$$M^{(2,2,1)} \cong S^{(2,2,1)} \oplus S^{(3,1,1)} \oplus 2S^{(3,2)} \oplus 2S^{(4,1)} \oplus S^{(5)}.$$

Note that Proposition 32 is a consequence of Young's rule. We shall end with a couple of examples illustrating Young's rule.

Example 34. Note that $K_{(n)\mu} = 1$ as there is only one (n)-semistandard tableau of content μ, formed by filling in all the required entries in order. Then Young's Rule implies that every M^μ contains exactly one copy of the trivial representation $S^{(n)}$ (see Example 17).

Example 35. Since a semistandard tableau with content (1^n) is just a standard tableau, we have $K_{\lambda(1^n)} = f^\lambda$ (the number of standard λ-tableaux). So Young's rule says that $M^{(1^n)} \cong \bigoplus_\lambda f^\lambda S^\lambda$. But from Example 8 we saw that $M^{(1^n)}$ is simply the regular representation. By taking the magnitude of the characters of both sides, we get another proof of the identity $n! = \sum_{\lambda \vdash n} (f^\lambda)^2$ that we saw in Theorem 23.

References

[Fu] W. Fulton: *Young Tableaux: With Applications to Representation Theory and Geometry*. Cambridge: Cambridge University Press 1997.

[FH] W. Fulton and J. Harris: *Representation Theory: A First Course*. New York: Springer 1991 (GTM **129**).

[Sa] B. E. Sagan: *The Symmetric Group: Representations, Combinatorial Algorithms, and Symmetric Functions*. New York: Springer 2001 (GTM **203**).

[Se] J-P. Serre: *Linear Representations of Finite Groups*. New York: Springer 1977 (GTM **42**).

[VO] A. M. Vershik and A. Yu. Okounkov: A new approach to the representation theory of symmetric groups. II, *Zapiski Nauchn. Semin. POMI* **307** (2004), 57–98. English transl.: *J. Math. Sci. (New York)* **131** #2 (2005), 5471–5494.

FACULTY FEATURE ARTICLE

5

Arrow's Impossibility Theorem: Two Simple Single-Profile Versions

Allan M. Feldman[†]
Department of Economics
Brown University
Providence, RI 02912
Allan_Feldman@Brown.edu
http://www.econ.brown.edu/fac/allan_feldman

Roberto Serrano[‡]
Department of Economics
Brown University
Providence, RI 02912
IMDEA-Social Science
Madrid, Spain
Roberto_Serrano@Brown.edu
http://www.econ.brown.edu/faculty/serrano

Abstract

In this paper we provide two simple new versions of Arrow's impossibility theorem, in a model with only one preference profile. Both versions are transparent, requiring minimal mathematical sophistication. The first version assumes there are only two people in society, whose preferences are being aggregated; the second version assumes two or more people. Both theorems rely on assumptions about diversity of preferences, and we explore alternative notions of diversity at some length. Our first theorem also uses a neutrality assumption, commonly used in the literature; our second theorem uses a neutrality/monotonicity assumption, which is stronger and less commonly used. We provide examples to illustrate our points.

5.1 Introduction

In 1950 Kenneth Arrow ([Ar1],[Ar2]) provided a striking answer to a basic abstract problem of democracy: how can the preferences of many individuals be aggregated into social preferences? The starkly negative answer, known as Arrow's impossibility theorem, was that every conceivable aggregation method has some flaw. That is, a handful of reasonable-looking axioms, which one thinks an aggregation procedure should satisfy, lead to impossibility: the axioms are mutually inconsistent. This impossibility theorem created a large literature and major field called social choice theory; see for example, Suzumura's ([Su]) Introduction to the *Handbook of Social Choice and Welfare*, and the Campbell and Kelly ([CK]) survey in the same volume.[1]

[†]Allan M. Feldman was born and grew up in New Jersey. He received an Sc.B. degree in mathematics from the University of Chicago and a Ph.D. in economics from Johns Hopkins University. He is a professor of economics at Brown University and has taught at Brown since 1971.

[‡]Roberto Serrano was born and grew up in Madrid, Spain. He received an A.B. in economics from Universidad Complutense de Madrid and a Ph.D. in economics from Harvard University. He has been at Brown since 1992, where he is now the Harrison S. Kravis University Professor of Economics. He is also a Research Associate in IMDEA (Madrid Institute for Advanced Studies).

[1]The theorem has also had a major impact on the larger fields of economics and political science, as well as on distant fields like mathematical biology. (See, *e.g.*, Day and McMorris ([DM]).)

In this paper we develop two very simple versions of Arrow's impossibility theorem. Our models are so-called single-profile models. This means impossibility is demonstrated in the context of one fixed profile of preferences, rather than in the (standard) Arrow context of many varying preference profiles.

Single-profile Arrow theorems were first proved in the late 1970's and early 1980's by Parks ([Pa]), Hammond ([Ha]), Kemp and Ng ([KN]), Pollak ([Po]), Roberts ([Ro]) and Rubinstein ([Ru]). Single-profile theorems were developed in response to an argument of Paul Samuelson ([Sa1]) against Arrow. Samuelson claimed that Arrow's model, with varying preference profiles, is irrelevant to the classical problem of maximizing a Bergson-Samuelson-type social welfare function (Bergson ([Be])), which depends on a given set of ordinal utility functions, that is, a fixed preference profile. The single-profile Arrow theorems established that negative results, such as dictatorship, or illogic of social preferences, or, more generally, impossibility of aggregation, could be proved with one fixed preference profile (or set of ordinal utility functions), provided the profile is "diverse" enough.

This paper has two purposes. The first is to provide two short and transparent single-profile Arrow theorems. In addition to being short and simple, our theorems do not require the existence of large numbers of alternatives. Our second purpose is to explore the meaning of preference profile diversity. Our first Arrow impossibility theorem, which is extremely easy to prove, assumes that there are only two people in society. The proof relies on a neutrality assumption and our first version of preference diversity, which we call simple diversity. In our second Arrow impossibility theorem, which is close to Pollak's ([Po]) version, there are two or more people. For this version we strengthen neutrality to neutrality/monotonicity, and we use a second, stronger version of preference diversity, which we call complex diversity.

Other recent related literature includes Geanakoplos ([Ge]), who has three very elegant proofs of Arrow's theorem in the standard multi-profile context, and Ubeda ([Ub]) who has another elegant multi-profile proof.[2] These proofs, while short, are mathematically much more challenging than ours. Reny ([Re]) has an interesting side-by-side pair of (multi-profile) proofs, of Arrow's theorem and the related theorem of Gibbard and Satterthwaite.

5.2 The Model

We assume a society with two or more individuals, and three or more alternatives. A specification of the preferences of all individuals is called a preference profile. In our theorems there is only one preference profile. The preference profile is somehow transformed into a social preference relation. This might be done through a voting process, through the actions of an enlightened government, or by the force of a dictator. Any kind of social choice process is possible in Arrow's world. The individual preference relations are all assumed to be complete and transitive. Both the individual and the social preference relations allow indifference. The following notation is used: Generic alternatives are $x, y, z, w, \ldots$. Particular alternatives are $a, b, c, d, \ldots$. A generic person is labeled $i, j, k, \ldots$; a particular person is $1, 2, 3, \ldots$. Person i's preference relation is R_i. xR_iy means person i prefers x to y or is indifferent between them; xP_iy means i prefers x to y; xI_iy means i is indifferent between them. Society's preference relation is R. xRy means society prefers x to y or is indifferent between them; xPy means society prefers x to y; xIy means society is indifferent between them. We start with the following assumptions[3]:

(1) **Complete and transitive social preferences**. The social preference relation R is complete and transitive.

(2.a) **Weak Pareto principle**. For all x and y, if xP_iy for all i, then xPy.

[2] Ubeda also emphasizes the importance of (multi-profile) neutrality, similar to but stronger than the assumption we use in this paper, and much stronger than Arrow's independence assumption, and he provides several theorems establishing neutrality's equivalence to other intuitively appealing principles.

[3] Assumptions are just assumptions, and are not necessarily true. In fact, Arrow's problem is to show that a set of assumptions is inconsistent: if all but one are true, then the remaining one must be false.

(2.b) **Strong Pareto principle**. For all x and y, if xR_iy for all i, and xP_iy for some i, then xPy.

(3.a) **Neutrality**. Suppose individual preferences for w versus z are identical to individual preferences for x versus y. Then the social preference for w versus z must be identical to the social preference for x versus y. Formally: For all x, y, z, and w, assume that, for all i, xP_iy if and only if wP_iz and zP_iw if and only if yP_ix. Then wRz if and only if xRy, and zRw if and only if yRx.

(4) **No dictator**. There is no dictator. Individual i is a **dictator** if, for all x and y, xP_iy implies xPy.

(5.a) **Simple diversity**. There exists a triple of alternatives x, y, z such that xP_iy for all i, but opinions are split on x versus z and on y versus z. That is, some people prefer x to z and some people prefer z to x, and, similarly, some people prefer y to z and some people prefer z to y.

Note that we have two alternative versions of the Pareto principle here. The first (weak Pareto) is more common in the Arrow's theorem literature (e.g., see Campbell and Kelly ([CK, p. 42])). We will use the strong Pareto principle in our two-person impossibility theorem below, and the weak Pareto principle in our two-or-more person impossibility theorem. Neutrality, assumption (3.a), and simple diversity, assumption (5.a), are so numbered because we will introduce alternatives later.

Also note that the no dictator assumption is different in a world with a single preference profile from what it is in the multi-profile world. For example, in the single-profile world, if all individuals have the same preferences, and if Pareto holds (weak or strong), then by definition everyone is a dictator. Or, if individual i is indifferent among all the alternatives, he is by definition a dictator. We will discuss this possibility of innocuous dictatorship in Section 5.9 below.

5.3 Some Examples in a Two-Person Model

We illustrate with a few simple examples. For these there are two people and three alternatives, and we assume no individual indifference between any pair of alternatives. Given that we aren't allowing individual indifference, the two Pareto principles collapse into one. Preferences of the two people are shown by listing the alternatives from top (most preferred) to bottom (least preferred). In our examples, the last column of the table shows what is being assumed about society's preferences. The comment below each example indicates which desired property is breaking down. The point of these examples is that if we are willing to discard any one of our five basic assumptions, the remaining four may be mutually consistent.

Person 1	Person 2	Society (Majority Rule)
a	c	
b	a	aPb, aIc, & bIc
c	b	

Example 1. Transitivity for social preferences fails. Transitivity for R implies transitivity for I. This means aIc & cIb should imply aIb. But we have aPb.

Person 1	Person 2	Society
a	c	
b	a	$aIbIc$
c	b	

Example 2. Pareto (weak or strong) fails, because aP_1b and aP_2b should imply aPb. But we have aIb.

Person 1	Person 2	Society
a	c	a
b	a	c
c	b	b

Example 3. Neutrality fails. Compare the social treatment of a versus c, where the two people are split and person 1 gets his way, to the social treatment of b versus c, where the two people are split and person 2 gets his way.

Person 1	Person 2	Society (1 is Dictator)
a	c	a
b	a	b
c	b	c

Example 4. There is a dictator.

Note that Examples 1 through 4 all use the same profile of individual proferences, which satisfies the simple diversity assumption. The next example modifies the individual preferences:

Person 1	Person 2	Society (Majority Rule)
a	c	
c	a	aIc
b	b	aPb & cPb

Example 5. Simple diversity fails. Opinions are no longer split over two pairs of alternatives.

5.4 Neutrality, Independence, and Some Preliminary Arrow Paradoxes

One of the most controversial of Arrow's original assumptions was independence of irrelevant alternatives. We did not define it above because it does not play a direct role in single-profile Arrow theorems; however it lurks behind the scenes. Therefore we define it at this point. Independence requires the existence of multiple preference profiles, and to accommodate multiple profiles, we use primes: Person i's preference relation was shown as R_i above, and society's as R; at this point we will write R_i' and R' for alternative preferences for person i and society, respectively. Now consider a pair of alternatives x and y. Arrow's independence of irrelevant alternatives condition requires appropriate consistency in the social ranking of x and y as individual preferences switch from unprimed to primed. More formally:

(6) **Independence**. Let $R_1, R_2, \ldots$ and R be one set of individual and social preference relations and $R_1', R_2', \ldots$ and R' be another. Assume that for all i, xP_iy if and only if $xP_i'y$ and $yP_i'x$ if and only if yP_ix. Then xRy if and only if $xR'y$ and $yR'x$ if and only if yRx.

Note the parallel between the independence assumption and the neutrality assumption. Independence involves multiple preference profiles whereas our version of neutrality assumes there is one preference profile. Independence focuses on a pair of alternatives and switches between two preference profiles, one unprimed and the other primed. It says that if the x versus y individual preferences are the same under the two preference profiles, then the x versus y social preference must also be the same. This statement is of course meaningless if there is only one preference profile. The closest analogy when there is only one preference profile is neutrality, which says that if individual preferences regarding x versus y under the one fixed preference profile are the same as individual preferences regarding w versus z under that profile, then the x versus y social preference must be the same as the w versus z social preference.

In short, in a single-profile model, independence is a vacuous assumption, and its natural replacement is neutrality.[4]

This natural replacement, however, prompted Samuelson to launch an attack in [Sa2] directed at the Kemp's and Ng's neutrality assumption in [KN]. Samuelson called neutrality, among other things, "anything but reasonable," and "gratuitous." ([Sa2]) He offered the following *reductio ad absurdum* example:

[4]The definition of neutrality can be easily extended to a multi-profile model, and neutrality is a stronger assumption than independence in such a model.

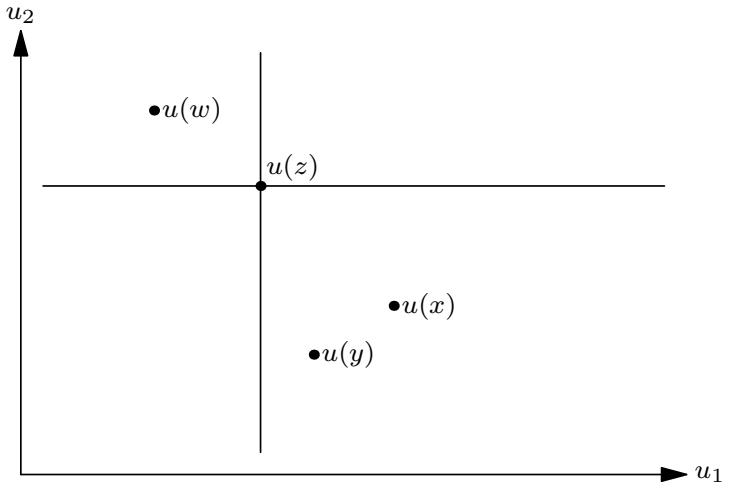


Figure 5.1: Fleurbaey and Mongin's Arrow impossibility argument.

Example 6 (Samuelson's Chocolates)**.** There are two people. There is a box of 100 (indivisible) chocolates to be distributed between them. They both like chocolates, and each is hungry enough to eat them all. The alternatives are $x_0 = (100, 0), x_1 = (99, 1), x_2 = (98, 2), \ldots$, where the first number is the number of chocolates going to person 1 and the second is the number going to person 2.

Many ethical observers, looking at this society, would say that x_1 is better than x_0. That is, $x_1 P x_0$. That is, it would be a good thing to take a chocolate from person 1, when he has 100 of them, and give it to person 2. Note that $x_0 P_1 x_1$ and $x_1 P_2 x_0$.

Now consider any $k < 100$. The individual preferences are $x_k P_1 x_{100}$ and $x_{100} P_2 x_k$, similar to the individual preferences for x_0 versus x_1. By neutrality, $x_{100} P x_k$! That is, society should give all the chocolates to person 2!

Samuelson's chocolates example is a vivid attack on neutrality, but should not be viewed as a compelling reason to drop it. One response to the example is to say society should not decide that x_1 is better than x_0 in the first place; if society simply found x_0 and x_1 equally good (contrary to the instincts of the chocolate redistributionist), neutrality would have implied that all the x's are socially indifferent. This would have been perfectly logical. Another response is to observe that neutrality is a property of extremely important and widely used decision-making procedures, particularly majority voting, and therefore cannot be lightly dismissed. In fact, any social decision procedure that simply counts instances of xP_iy, yP_ix, and xI_iy, but does not weigh strength of feelings, satisfies neutrality.

Samuelson ([Sa2]) also offered a graphical argument against Arrow's theorem with neutrality, an argument that was simplified and improved years later by Fleurbaey and Mongin ([FM]), as follows:

Fleurbaey and Mongin Graphical Arrow Impossibility Argument. Assume that there are two people, and some set of alternatives $x, y, z, \ldots$. Assume the individuals have utility functions u_1 and u_2, so $u_1(x)$, for example, represents person 1's utility level from alternative x.

Consider the graph in Figure 5.1. Utility levels of individuals 1 and 2 are on the horizontal and vertical axes, respectively. Each alternative shows up in the graph as a utility pair, for instance $u(z) = (u_1(z), u_2(z))$ represents alternative z. We start at $u(z)$ and draw horizontal and vertical lines through it, creating four quadrants.

Now assume complete and transitive social preferences, strong Pareto and neutrality. Take two alternatives, say x and y, whose utility vectors are within the *southeast* quadrant. Choose them so that $u(x)$ is northeast of $u(y)$.

Social indifference between z and x is impossible, for the following reasons: First, by neutrality, if zIx, then zIy, must also hold. Second, if zIx and zIy, then xIy by transitivity. But third,

since $u(x)$ is northeast of $u(y)$, xPy by Pareto.

Therefore either society prefers z to x, or society prefers x to z. Suppose xPz. Now consider another alternative w. By neutrality, if $u(w)$ is in the *northwest* quadrant (as in Figure 5.1), xPz implies zPw. By neutrality, if $u(w)$ is in the *southeast* quadrant, xPz implies wPz. By strong Pareto, if $u(w)$ is in the *northeast* quadrant, wPz. By strong Pareto, if $u(w)$ is in the *southwest* quadrant, zPw. But this argument establishes that social preferences (for w versus z) are always exactly the same as person 1's; that is, person 1 is a dictator. Had we started out by assuming zPx, person 2 would have been the dictator. In short, the graph produces an Arrow impossibility. □

There are two drawbacks to the Fleurbaey/Mongin/(Samuelson) graphical impossibility argument. First, it has the disadvantage that it requires the use of the utility functions u_1 and u_2—it is cleaner to dispense with utility functions and simply use preference relations for individuals. Second, it incorporates a crucial diversity assumption without being explicit about it. Assuming the existence of the triple of utility vectors $u(x)$, $u(y)$, and $u(z)$, with their respective locations in the utility diagram, is in fact exactly the assumption of simple diversity: both 1 and 2 prefer x to y, but opinions are split on x versus z and opinions are split on y versus z. In Theorem 8 below, we make this assumption explicit.

5.5 Arrow Impossibility Theorem, $n = 2$

We are ready to turn to our own simple version of Arrow's impossibility theorem, in the single-profile model. Throughout this section, we assume there are two people in society. We will show that our five assumptions, complete and transitive social preferences, strong Pareto, neutrality, simple diversity, and no dictator, are mutually inconsistent.

First we establish Proposition 7, which is by itself a very strong result. This proposition corresponds to the Samuelson's chocolates example, and so we call it Samuelson's chocolates proposition. Then we prove our first simple version of Arrow's theorem.

Proposition 7 (Samuelson's Chocolates)**.** *Assume $n = 2$. Assume the strong Pareto principle and neutrality. Suppose for some pair of alternatives x and y, and for the two people i and j, xP_iy and yP_jx. Suppose that xPy. Then person i is a dictator.*

Proof. Let w and z be any pair of alternatives. Assume wP_iz. We need to show that wPz must hold. If wR_jz, then wPz by strong Pareto. If not, wR_jz, then zP_jw by completeness for j's preference relation, and then wPz by neutrality. □

Theorem 8 (Arrow Impossibility Theorem)**.** *Assume $n = 2$. The assumptions of complete and transitive social preferences, strong Pareto, neutrality, simple diversity, and no dictator are mutally inconsistent.*

Proof. By simple diversity there exist x, y and z such that xP_iy for $i = 1, 2$, but such that opinions are split on x versus z, and on y versus z.

Now xPy by the Pareto principle, weak or strong. Since opinions are split on x versus z, one person prefers x to z, while the other prefers z to x. If xPz, then the person who prefers x to z is a dictator by Proposition 7. If zPx, then the person who prefers z to x is a dictator by Proposition 7.

Suppose then that xIz. Then zIx. By transitivity, zIx and xPy implies zPy. But opinions are split on y versus z. Therefore one person prefers z to y, and the other person prefers y to z. By Proposition 7, the person who prefers z to y is a dictator. We have shown that whatever the social preference for x and z might be, there must be a dictator. □

5.6 Trying to Generalize to an n-Person Model

In what follows we seek to generalize our version of Arrow's theorem to societies with two or more people. In order to get an impossibility theorem when $n \geq 2$, we need to strengthen some of our basic assumptions. We start with the neutrality assumption. We will strengthen it to a single-profile version of what is called neutrality/monotonicity.[5] The intuition is that if everybody who prefers

[5]See Blau & Deb ([BD]), who call the multi-profile analog "full neutrality and monotonicity"; Sen ([Se]), who calls it NIM; and Pollak ([Po]), who calls it "nonnegative responsiveness."

x over y also prefers z over w, and everybody who prefers w over z also prefers y over x, then if society prefers x to y, it should also prefer w to z.

(3.b) **Neutrality/monotonicity**. For all x, y, z, and w, assume that for all i, xP_iy implies wP_iz, and that for all i, zP_iw implies yP_ix. Then xPy implies wPz.

This strengthening of the neutrality assumption does not, by itself, give us an Arrow impossibility theorem when there are two or more people. In Example 9 below there are three people and four alternatives, a, b, c and d. The preferences of individuals 1, 2, and 3 are shown in the first 3 columns of the table. The fourth column shows social preferences under majority rule, which is used here, as in Examples 1 and 5, to generate the social preference relation.

Person 1	Person 2	Person 3	Society (Majority Rule)
a	c	a	a
b	a	c	c
c	b	d	b
d	d	b	d

Example 9. None. The complete and transitive social preferences assumption is satisfied, as are Pareto, neutrality/monotonicity, simple diversity, and no dictator. Majority rule works fine. There is no Arrow impossibility.

Example 9 shows that when $n \geq 2$ there is no Arrow impossibility, under the assumptions of complete and transitive social preferences, Pareto, neutrality/monotonicity, simple diversity, and no dictator.

5.7 Diversity

In this section we will modify the diverse preferences assumption.

Before doing so, let's revisit the assumption in the two-person world. In that world, simple diversity says there must exist a triple of alternatives x, y, z, such that xP_iy for $i = 1, 2$, but such that opinions are split on x versus z and on y versus z. That is, one person prefers x to z, while the other prefers z to x, and one person prefers y to z, while the other prefers z to y. Given our assumption that individual preferences are transitive, it must be the case that the two people's preferences over the triple can be represented as follows:

Person i	Person j
x	z
y	x
z	y

Table 5.1: Simple diversity array, $n = 2$.

Note that this is exactly the preference profile pattern of Examples 1, 2, 3 and 4. [6]

A somewhat similar array was used by Arrow in the proof of his impossibility theorem.[7] For now assume that V is any non-empty set of people in society, that V^C is the complement of V, and that V can be partitioned into two non-empty subsets V_1 and V_2. (Note that V^C may be empty.) The standard Arrow preference array looks like this:

Now, let's return to the question of how to modify the diverse preferences assumption. Example 9 shows that we cannot stick with the simple diversity array and still get an impossibility result. We

[6] Readers familiar with social choice theory will recognize the simple diversity array as being two thirds of the Condorcet voting paradox array. Condorcet's array simply adds a third person, say k, who prefers y to z to x.

[7] The array to which we now turn has been used by Arrow ([Ar2, p. 58]) and by many others since, including us ([FS, p. 294]).

People in V_1	People in V_2	People in V^C
x	z	y
y	x	z
z	y	x

Table 5.2: Standard Arrow array.

might start with the Condorcet voting paradox array, but if $n \geq 4$, we would have to worry about the preferences of people other than i, j and k. That suggests using something like the standard Arrow array. However, assuming the existence of a triple x, y, and z, and preferences as per that array, for *every subset of people V and every partition of V*, is an unnecessarily strong diversity assumption.

An even stronger diversity assumption was in fact used by Parks ([Pa]), Pollak and other originators of single-profile Arrow theorems. Pollak ([Po]) is clearest in his definition. His condition of "unrestricted domain over triples" requires the following: Imagine "any logically possible sub-profile" of individual preferences over three "hypothetical" alternatives x, y and z. Then there exist three actual alternatives a, b and c for which the sub-profile of preferences exactly matches that "logically possible sub-profile" over x, y and z. We will call this **Pollak diversity**. Let us consider what this assumption requires in the simple world of strict preferences, two people, and three alternatives. Pollak diversity would require that every one of the following arrays be represented, somewhere in the actual preference profile of the two people over the actual alternatives:

1	2	1	2	1	2	1	2	1	2	1	2
x	x	x	x	x	y	x	y	x	z	x	z
y	y	y	z	y	x	y	z	y	x	y	y
z	z	z	y	z	z	z	x	z	y	z	x

Table 5.3: Pollak diversity arrays, $n = 2$.

Note that the number of arrays in the table above is $3! = 6$. If n were equal to 3 we would have triples of columns instead of pairs, and there would have to be $(3!)^2 = 36$ such triples. With n people, the number of required n-tuples would be $(3!)^{n-1}$. In short, the number of arrays required for Pollak diversity rises exponentially with n. The number of alternatives rises with the number of required arrays, although not as fast because of array overlaps. Parks ([Pa]) uses an assumption ("diversity in society") that is very similar to Pollak's, although not so clear, and he indicates that it "requires at least 3^n alternatives…".

Pollak diversity is actually much stronger than necessary. We will weaken it as follows. We will not assume the existence of a triple x, y and z and every conceivable array of preferences on that triple. Nor will we assume the existence of a triple x, y and z and every conceivable array of preferences on that triple, but restricted to sets V, V_1, V_2, and V^C, as per the description of the standard Arrow array. Rather, we will simply assume the existence of triple x, y and z, and the standard Arrow array preferences on that triple, when it really matters. For our purposes, it really matters when the set V referenced in the description of the standard Arrow array is a decisive set. This is defined as follows:

Definition 10. We say that a set of people V is **decisive** if it is non-empty and if, for all alternatives x and y, if xP_iy for all i in V, then xPy.

It is appropriate to make a few comments about the notion of decisiveness. First, note that if person i is a dictator, then i by himself is a decisive set, and any set containing i is also decisive. Also, note that the Pareto principle (weak or strong) implies the set of all people is decisive. Second, in a multi-preference profile world, decisiveness for V would be a far stronger assumption that it is in the single-profile world, since it would require that (the same) V prevail no matter how

preferences might change. We only require that V always prevail under the single preference profile.

Our diversity assumption is now modified as follows:

(5.b) **Complex diversity**. For any decisive set V with 2 or more members, there exists a triple of alternatives x, y, z, such that xP_iy for all i in V; such that yP_iz and zP_ix for everyone outside of V; and such that V can be partitioned into non-empty subsets V_1 and V_2, where the members of V_1 all put z last in their rankings over the triple, and the members of V_2 all put z first in their rankings over the triple.

The assumption of complex diversity means that for any decisive set V with two or more members, there is a triple x, y, and z, and a partition of V, which produces exactly the standard Arrow array shown above.

Simple diversity and complex diversity are related in the following way: If $n = 2$ and weak Pareto holds, they are equivalent. If $n > 2$, neither one implies the other, but they are both implied by Pollack diversity.

Referring back to Example 9 of the previous section, consider persons 2 and 3. Under simple majority rule, which was assumed in the example, they constitute a decisive coalition. However the complex diversity assumption fails in the example, because there is no way to define the triple x, y, z so as to get the standard Arrow array, when $V = \{2, 3\}$. Therefore complex diversity rules out that example.

Example 11 below modifies Example 9 so that, for the decisive set $V = \{2, 3\}$, the preference profile is consistent with complex diversity. (This example is created from Example 9 by switching alternatives a and b in person 3's ranking. Let $V_1 = \{2\}$, $V_2 = \{3\}$, and $V^C = \{1\}$. The triple x, y, z is now c, a, b.) Now that preferences have been modified consistent with our new diversity assumption, an Arrow impossibility pops up.

Person 1	Person 2	Person 3	Society (Majority Rule)
a	c	b	
b	a	c	aPb, bPc, cPa
c	b	d	aPd, bPd, cPd
d	d	a	

Example 11. Transitivity for social preference fails with a strict social preference cycle among a, b, and c. Society prefers a to b, b to c, and, irrationally, c to a.

Example 11 could be further modified by dropping alternative d, in which case it would become the Condorcet voting paradox array. (See footnote 6 above.) It would then have three people and three alternatives, and would satisfy complex diversity. Recall that Pollack diversity in the three-person case would require at least 36 n-tuples of alternatives, and that Parks diversity would require at least $3^n = 27$ alternatives. The point here is that that complex diversity is a much less demanding assumption, and requires many fewer alternatives, than Pollack diversity.

Complex diversity captures the idea of moderately divergent opinions when there are three or more people in society. It requires that when V is a decisive set with two or more members, there must exist some triple of alternatives x, y, and z about which there is basic disagreement, both within V (with those in V_1 putting z at the bottom and those in V_2 putting z and the top), and between V and V^C (with those in V preferring x over y, while those in V^C preferring y over x). But it is not an overly strong assumption, like Pollak diversity, nor does it require an enormous number of alternatives. We do not claim that the complex diversity assumption has the moral appeal of the Pareto principal or the no dictatorship assumption, but it is a plausible possibility, and one can very easily imagine real examples of preferences like those assumed in Example 11 above.

We will finish this discussion of diversity by noting our complex diversity assumption might be modified in either of two directions: It could be strengthened, by dropping the requirement in the definition that V be a decisive set. We will call the diversity assumption so modified **arbitrary V complex diversity**. This assumption would be closer to Pollak diversity. Alternatively, the complex diversity assumption could be weakened, by adding the requirement that V be a decisive

set of minimal size. We will call the diversity assumption so modified **minimally-sized decisive V complex diversity**. We will briefly refer to both of these modifications at the end of the next section.

5.8 Arrow/Pollak Impossibility, $n \geq 2$

We now proceed to a proof of our second single-profile Arrow's theorem, which, unlike Theorem 8, is not restricted to a two-person society.[8] Although Pollak made a much stronger diversity assumption than we use, and although Parks ([Pa]), Hammond ([Ha]), and Kemp and Ng ([KN]), preceded Pollak with single-profile Arrow theorems, we will call this the Arrow/Pollak Impossibility Theorem, because of the similarity of our proof to his. But first, we need a proposition paralleling Proposition 7:

Proposition 12. *Assume $n \geq 2$ and neutrality/monotonicity. Assume there is a non-empty group of people V and a pair of alternatives x and y, such that xP_iy for all i in V and yP_ix for all i not in V. Suppose that xPy. Then V is decisive.*

Proof. This follows immediately from neutrality/monotonicity. □

Theorem 13 (Arrow/Pollak Impossibility Theorem)**.** *Assume $n \geq 2$. The assumptions of complete and transitive social preferences, weak Pareto, neutrality/monotonicity, complex diversity, and no dictator are mutually inconsistent.*

Proof. By the weak Pareto principle, the set of all individuals is decisive. Therefore decisive sets exist. Let V be a decisive set of minimal size, that is, a decisive set with no proper subsets that are also decisive. We will show that there is only one person in V, which will make that person a dictator. This will establish Arrow's theorem.

Suppose to the contrary that V has 2 or more members. By the complex diversity assumption there is a triple of alternatives x, y, and z, and a partition of V into non-empty subsets V_1 and V_2, giving the standard Arrow array as shown above. Since V is decisive, it must be true that xPy. Next we consider the social preference for x versus z.

Case 1. Suppose zRx. Then zPy by transitivity. Then V_2 becomes decisive by Proposition 12 above. But this is a contradiction, since we assumed that V was a decisive set of minimal size.

Case 2. Suppose not zRx. Then the social preference must be xPz, by completeness. But in this case V_1 is getting its way in the face of opposition by everone else, and by Proposition 12 above V_1 is decisive, another contradiction. □

In Section 5.7 above we mentioned two alternative versions of complex diversity, a stronger version, arbitrary V complex diversity, and a weaker version, minimally-sized decisive V complex diversity. Either of these could be substituted for complex diversity in Theorem 13 above, without affecting the proof. Moreover, using the minimally-sized V complex diversity assumption would give the following near-converse to Arrow's theorem: If there is a dictator, then the minimally-sized V complex diversity assumption is satisfied. This follows immediately from the definition of minimally-sized V complex diversity. For if i is a dictator, then $\{i\}$ is a decisive set; so any minimally-sized decisive set can have only one member, and therefore cannot be partitioned into two non-empty subsets. Consequently the definition of minimally-sized V complex diversity is vacuously satisfied.

5.9 Innocuous Dictators

In the standard multi-profile world, where all preference profiles are allowed (the so-called "universality," or "full domain" assumption) a dictator is a very bad thing indeed. A dictator in such a world forces his (strict) preference for x over y even if everyone else prefers y over x. In our single-profile world, on the other hand, a dictator may be innocuous. For instance, if person i is indifferent between all pairs of alternatives, he is by definition a dictator, although a completely benign one. Or, if everyone has exactly the same preferences over the alternatives, and weak Pareto

[8] There is a similar proof, but for a multi-profile Arrow's theorem, in Feldman & Serrano ([FS]).

is satisfied, then each person is a dictator. Or, if in a committee of five people, three have identical preferences, and if they use majority rule, then the three with identical preferences are all dictators. (Note however that in a standard median voter model, the median voter is not necessarily a dictator. While his favorite alternative may be the choice of the committee, the committee's preferences over all pairs of alternatives will not necessarily agree with his preferences over those pairs of alternatives.)

Therefore we need to make a final comment about why dictatorship should worry us, even though some dictators are innocuous: While we assume a single-profile world in this paper, and while for certain given profiles dictatorship doesn't look bad, we must remember that there can be other single-profile worlds with different given preference profiles. So, while in some cases an innocuous dictatorship is acceptable, in many other cases it is very much unacceptable. Moreover, both of our diversity assumptions exclude vacuous dictatorship cases like the one in which all individuals have exactly the same preferences. In sum, even though single-profile analysis may permit innocuous dictators, dictatorship remains a very bad thing, and Arrow's theorem remains important.

5.10 Conclusions

We have presented two new single-profile Arrow impossibility theorems which are simple and transparent. Theorem 8, which requires that there are only two people, relies on a very simple and modest assumption about diversity of preferences within the given preference profile, and on a relatively modest neutrality assumption. Theorem 13, which allows for two or more people, uses a substantially more complex assumption about diversity of preferences within the given profile, and uses a stronger neutrality/monotonicity assumption. Both theorems establish that Arrow impossibility happens even if individual preferences about alternatives are given and fixed.

5.11 Acknowledgments

We want to thank Kenneth Arrow for having suggested the questions that led us to write this paper. We also thank Scott Kominers, our editor at The HCMR, for his careful reading and detailed comments. Finally, the discussion of Arrow's theorem made by Don Saari at the Brown Symposium for Undergraduates in the Mathematical Sciences was illuminating. Like him, we believe that this topic should be most intriguing to young mathematical minds.

References

[Ar1] Kenneth Arrow: A Difficulty in the Concept of Social Welfare, *J. of Polit. Econ.* **58** (1950), 328–346.

[Ar2] Kenneth Arrow: *Social Choice and Individual Values*, 2nd ed. New York: John Wiley & Sons 1963.

[BD] Julian Blau and Rajat Deb: Social Decision Functions and the Veto, *Econometrica* **45** (1977), 871–879.

[Be] Abram Bergson: A Reformulation of Certain Aspects of Welfare Economics, *Quar. J. of Econ.* **52** (1938), 310–334.

[CK] Donald Campbell and Jerry Kelly: Impossibility Theorems in the Arrovian Framework, in Kenneth Arrow, Amartya Sen, and Kotaro Suzumura: *Handbook of Social Choice and Welfare*, vol. 1. Amsterdam: Elsevier Science 2002.

[DM] William Day and F. R. McMorris: Axiomatic Consensus Theory in Group Choice and Biomathematics, *SIAM* Philadelphia (2003)

[FS] Allan Feldman and Roberto Serrano: *Welfare Economics and Social Choice Theory*, 2nd ed. New York: Springer 2006.

[FM] Marc Fleurbaey and Philippe Mongin: The news of the death of welfare economics is greatly exaggerated, *Social Choice and Welfare* **25** (2005), 381–418.

[Ge] John Geanakoplos: Three Brief Proofs of Arrow's Impossibility Theorem, *Econ. Theory* **26** (2005), 211–215.

[Ha] Peter Hammond: Why Ethical Measures of Inequality Need Interpersonal Comparisons, *Theory and Decision* **7** (1976), 263–274.

[KN] Murray Kemp and Yew-Kwang Ng: On the Existence of Social Welfare Functions, Social Orderings and Social Decision Functions, *Economica* **43** (1976), 59–66.

[Pa] Robert Parks: An Impossibility Theorem for Fixed Preferences: A Dictatorial Bergson-Samuelson Welfare Function, *Review of Econ. Stud.* **43** (1976), 447–450.

[Po] Robert Pollak: Bergson-Samuelson Social Welfare Functions and the Theory of Social Choice, *Quar. J. of Econ.* **93** (1979), 73–90.

[Re] Philip Reny: Arrow's Theorem and the Gibbard-Satterthwaite Theorem: a Unified Approach, *Econ. Letters* **70** (2001), 99–105.

[Ro] Kevin Roberts: Social Choice Theory: The Single-profile and Multi-profile Approaches, *Review of Econ. Stud.* **47** (1980), 441–450.

[Ru] Ariel Rubinstein: The Single Profile Analogues to Multi-Profile Theorems: Mathematical Logic's Approach, *International Econ. Review* **25** (1984), 719–730.

[Sa1] Paul Samuelson: Arrow's Mathematical Politics, in S. Hook, ed.: *Human Values and Economics Policy*. New York: New York University Press 1967, 41–52.

[Sa2] Paul Samuelson: Reaffirming the Existence of 'Reasonable' Bergson-Samuelson Social Welfare Functions, *Economica* **44** (1977), 81–88.

[Se] Amartya Sen: Social Choice Theory: A Re-Examination, *Econometrica* **45** (1977), 53–89.

[Su] Kotaro Suzumura: Introduction, in Kenneth Arrow, Amartya Sen, and Kotaro Suzumura, eds.: *Handbook of Social Choice and Welfare*, vol. 1. Amsterdam: Elsevier Science 2002.

[Ub] Luis Ubeda: Neutrality in Arrow and Other Impossibility Theorems, *Econ. Theory* **23** (2004), 195–204.

FACULTY FEATURE ARTICLE

6 The Congruent Number Problem

Keith Conrad[†]
University of Connecticut
Storrs, CT 06269
kconrad@math.uconn.edu

Abstract

We discuss a famous problem about right triangles with rational side lengths. This elementary-sounding problem is still not completely solved; the last remaining step involves the Birch and Swinnerton-Dyer conjecture, which is one of the most important open problems in number theory (right up there with the Riemann hypothesis).

6.1 Introduction

A right triangle is called **rational** when its legs and hypotenuse are all rational numbers. Examples of rational right triangles include Pythagorean triples like $(3, 4, 5)$. We can scale such triples to get other rational right triangles, like $(3/2, 2, 5/2)$. Of course, usually when two sides are rational the third side is not rational, such as in the $(1, 1, \sqrt{2})$ right triangle.

Any rational right triangle has a rational area, but not all (positive) rational numbers can occur as the area of a rational right triangle. For instance, no rational right triangle has area 1. This was proved by Fermat. The question we will examine here is: which rational numbers occur as the area of a rational right triangle?

Definition 1. A positive rational number n is called a **congruent number** if there is a rational right triangle with area n: there are rational $a, b, c > 0$ such that $a^2 + b^2 = c^2$ and $(1/2)ab = n$.

In Figure 6.1, there are rational right triangles with respective areas 5, 6, and 7, so these three numbers are congruent numbers.

This use of the word congruent has nothing to do (directly) with congruences in modular arithmetic. The etymology will be explained in Section 6.3. The history of congruent numbers can be found in [Di, Chap. XVI], where it is indicated that an Arab manuscript called the search for congruent numbers the "principal object of the theory of rational right triangles."

The congruent number problem asks for a description of all congruent numbers. Since scaling a triangle changes its area by a square factor, and every rational number can be multiplied by a suitable rational square to become a squarefree integer (*e.g.*, $18/7 = 3^2 \cdot 2/7$, so multiplying by $(7/3)^2$ produces the squarefree integer 14), we can focus our attention in the congruent number problem on squarefree positive integers. For instance, to say 1 is not a congruent number means no rational square is a congruent number.

When n is squarefree in $\mathbb{Z}^+$, we just need to find an integral right triangle whose area has squarefree part n to show n is a congruent number. Then writing the area as $m^2 n$ shows scaling the sides by m produces a rational right triangle with area n.

In Section 6.2, the parametrization of Pythagorean triples will be used to construct a lousy algorithm to generate all congruent numbers. The equivalence of the congruent number problem with a problem about rational squares in arithmetic progressions is in Section 6.3. Section 6.4 gives an equivalence between the congruent number problem and the search for rational points on $y^2 = x^3 - n^2 x$ where $y \neq 0$, which ultimately leads to a solution of the congruent number

[†]Keith Conrad received his undergraduate and graduate degrees in mathematics from Princeton (1992) and Harvard (1997). He became interested in number theory as a high school student at the Ross program at Ohio State University in 1986.

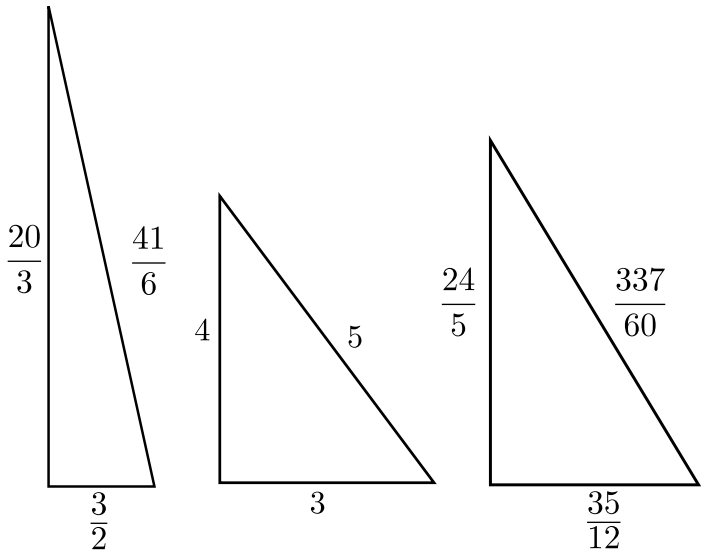


Figure 6.1: Rational right triangles with area 5, 6, and 7.

problem (depending in part on the Birch and Swinnerton-Dyer conjecture, a famous open problem in mathematics). In the appendices, we explain some algebraically mysterious formulas from our treatment using projective geometry and give a relation between the congruent number problem and other Diophantine equations.

6.2 A bad algorithm

There is a parametric formula for primitive Pythagorean triples and by using it we will make a small list of squarefree congruent numbers. Any primitive triple (with even second leg) is $(k^2 - \ell^2, 2k\ell, k^2 + \ell^2)$ where $k > \ell > 0$, $(k,\ell) = 1$, and $k \not\equiv \ell \bmod 2$. In Table 6.1 we list such primitive triples where $k + \ell \leq 9$. The squarefree part of the area is listed in the last column. Each number in the fourth column is a congruent number and each number in the fifth column is also a congruent number. The final row of the table explains how a rational right triangle with area 5 can be found.

k	ℓ	(a, b, c)	$(1/2)ab$	Squarefree part
2	1	$(3, 4, 5)$	6	6
4	1	$(15, 8, 17)$	60	15
3	2	$(5, 12, 13)$	30	30
6	1	$(35, 12, 37)$	210	210
5	2	$(21, 20, 29)$	210	210
4	3	$(7, 24, 25)$	84	21
8	1	$(63, 16, 65)$	504	126
7	2	$(42, 28, 53)$	630	70
5	4	$(9, 40, 41)$	180	5

Table 6.1: Congruent Numbers.

Notice 210 shows up twice in Table 6.1. Do other numbers which occur once also occur again? We will return to this question later.

Table 6.1 can be extended according to increasing values of $k + \ell$, and any squarefree congruent number eventually shows up in the last column, *e.g.*, the triangle $(175, 288, 337)$ with area $25200 = 7 \cdot 60^2$ occurs at $k = 16$ and $\ell = 9$. Alas, the table is *not* systematic in the appear-

ance of the last column: we can't tell by building the table when any particular number should occur, if at all, in the last column, so this method of generating (squarefree) congruent numbers is not a good algorithm. For instance, 53 is a congruent number, but it shows up for the first time when $k = 1873180325$ and $\ell = 1158313156$. (The corresponding right triangle has area $53 \cdot 297855654284978790^2$.)

Tabulations of congruent numbers can be found in Arab manuscripts from the 10th century, and 5 and 6 appear there. Fibonacci discovered in the 13th century that 7 is congruent and he stated that 1 is *not* congruent (that is, no rational right triangle has area equal to a perfect square). The first accepted proof is due to Fermat, who also showed 2 and 3 are not congruent numbers.

Theorem 2 (Fermat, 1640)**.** *The number* 1 *is not congruent.*

Proof. We will use the method of descent, which was discovered by Fermat on this very problem. Our argument is adapted from [Co, pp. 658–659].

Assume there is a rational right triangle with area 1. Calling the sides a/d, b/d, and c/d, where a, b, c, and d are positive integers, we have $a^2 + b^2 = c^2$ and $(1/2)ab = d^2$. (In other words, if there is a rational right triangle with area 1 then there is a Pythagorean triangle whose area is a perfect square. The converse is true as well.) Clearing the denominator in the second equation,

$$a^2 + b^2 = c^2, \quad ab = 2d^2. \tag{6.1}$$

We will show (6.1) has no positive integer solutions.

Assume there is a solution to (6.1) in positive integers. Let's show there is then a solution where a and b are relatively prime. Set $g = (a, b)$, so $g|a$ and $g|b$. Then $g^2|c^2$ and $g^2|2d^2$, so $g|c$ and $g|d$ (why?). Divide a, b, c, and d by g to get another 4-tuple of positive integers satisfying (6.1) with $(a, b) = 1$. So we may now focus on showing (6.1) has no solution in positive integers with the extra condition that $(a, b) = 1$.

We will do this using Fermat's method of descent: construct a new 4-tuple of positive integers a', b', c', d' satisfying (6.1) with $(a', b') = 1$ and $0 < c' < c$. Repeating this enough times, we reach a contradiction. Several times in the descent process we will use the following (or minor variations on it): two positive relatively prime integers whose product is a perfect square must each be perfect squares.

Now we start the descent. Since $ab = 2d^2$ and a and b are relatively prime, a or b is even but not both. Then $c^2 = a^2 + b^2$ is odd, so c is odd. Since ab is twice a square, $(a, b) = 1$, and a and b are positive, one is a square and the other is twice a square. The roles of a and b are symmetric, so without loss of generality a is even and b is odd. Then

$$a = 2k^2, \quad b = \ell^2$$

for some positive integers k and ℓ, with ℓ odd (because b is odd). The first equation in (6.1) now looks like $4k^4 + b^2 = c^2$, so $\frac{c+b}{2}\frac{c-b}{2} = k^4$. Because b and c are both odd and relatively prime, $(c + b)/2$ and $(c - b)/2$ are relatively prime. Therefore

$$\frac{c+b}{2} = r^4, \quad \frac{c-b}{2} = s^4$$

for some relatively prime positive integers r and s. Solve for b and c by adding and subtracting these equations:

$$b = r^4 - s^4, \quad c = r^4 + s^4,$$

so $\ell^2 = b = (r^2 + s^2)(r^2 - s^2)$. The factors $r^2 + s^2$ and $r^2 - s^2$ are relatively prime: any common factor would be odd (since ℓ is odd) and divides the sum $2r^2$ and the difference $2s^2$, so is a factor of $(r^2, s^2) = 1$. Since the product of $r^2 + s^2$ and $r^2 - s^2$ is an odd square and one of these is positive, the other is positive and

$$r^2 + s^2 = t^2, \quad r^2 - s^2 = u^2 \tag{6.2}$$

for odd positive integers t and u which are relatively prime. Since $u^2 \equiv 1 \bmod 4$, $r^2 - s^2 \equiv 1 \bmod 4$, which forces r to be odd and s to be even. Solving for r^2 in (6.2),

$$r^2 = \frac{t^2+u^2}{2} = \left(\frac{t+u}{2}\right)^2 + \left(\frac{t-u}{2}\right)^2, \tag{6.3}$$

where $(t \pm u)/2 \in \mathbb{Z}$ since t and u are odd.

Equation (6.3) will give us a "smaller" version of (6.1). Setting

$$a' = \frac{t+u}{2}, \ \ b' = \frac{t-u}{2}, \ \ c' = r,$$

we have $a'^2 + b'^2 = c'^2$. From $(t,u) = 1$ we get $(a',b') = 1$. Moreover, using (6.2), $a'b' = (t^2-u^2)/4 = 2s^2/4 = 2(s/2)^2$. Let $d' = s/2 \in \mathbb{Z}$, so we have a new solution (a',b',c',d') to (6.1). Since $0 < c' = r \leq r^4 < r^4 + s^4 = c$, by descent we get a contradiction. □

Theorem 2 leads to a weird proof that $\sqrt{2}$ is irrational. If $\sqrt{2}$ were rational then $\sqrt{2}$, $\sqrt{2}$, and 2 would be the sides of a rational right triangle with area 1. This is a contradiction of 1 not being a congruent number!

6.3 Relation to Arithmetic Progressions of Three Squares

The three squares 1, 25, and 49 form an arithmetic progression with common difference 24. The squarefree part of 24 is 6. This is related to 6 being a congruent number, by the following theorem.

Theorem 3. *Let $n > 0$. There is a one-to-one correspondence between right triangles with area n and 3-term arithmetic progressions of squares with common difference n: the sets*

$$\{(a,b,c) : a^2 + b^2 = c^2, (1/2)ab = n\}, \ \ \{(r,s,t) : s^2 - r^2 = n, t^2 - s^2 = n\}$$

are in one-to-one correspondence by

$$(a,b,c) \mapsto ((b-a)/2, c/2, (b+a)/2), \quad (r,s,t) \mapsto (t-r, t+r, 2s).$$

Proof. It is left to the reader to check the indicated functions take values in the indicated sets and that the correspondences are inverses of one another: if you start with an (a,b,c) and make an (r,s,t) from it, and then form an (a',b',c') from this (r,s,t), you get back the original (a,b,c). Similarly, starting with an (r,s,t), producing an (a,b,c) from it and then producing an (r',s',t') from that returns the same (r,s,t) you started with. □

How could the correspondence in Theorem 3 be discovered? When $s^2 - r^2 = n$ and $t^2 - s^2 = n$, adding gives $t^2 - r^2 = 2n$, so $(t-r)(t+r) = 2n$. This suggests using $a = t-r$ and $b = t+r$. Then $a^2 + b^2 = 2(t^2+r^2) = 2(2s^2) = (2s)^2$, so use $c = 2s$.

When $n > 0$ is rational, the correspondence in Theorem 3 preserves rationality and positivity/monotonicity: (a,b,c) is a rational triple if and only if (r,s,t) is a rational triple, and $0 < a < b < c$ if and only if $0 < r < s < t$. Therefore, n is congruent if and only if there is a rational square s^2 such that $s^2 - n$ and $s^2 + n$ are also squares. Note the correspondence in Theorem 3 involves not the squares in arithmetic progression but their square roots r, s, and t.

Example 4. For $n = 6$, using $(a,b,c) = (3,4,5)$ in Theorem 3 produces $(r,s,t) = (1/2, 5/2, 7/2)$, whose termwise squares are the arithmetic progression $1/4, 25/4, 49/4$ with common difference 6.

Example 5. Taking $n = 5$ and $(a,b,c) = (3/2, 20/3, 41/6)$, the correspondence in Theorem 3 yields $(r,s,t) = (31/12, 41/12, 49/12)$: the rational squares $(31/12)^2$, $(41/12)^2$, $(49/12)^2$ are an arithmetic progression with common difference 5.

Example 6. Since Fermat showed 1 and 2 are not congruent numbers, there is no arithmetic progression of 3 rational squares with common difference 1 or 2 (or, more generally, common difference a nonzero square or twice a nonzero square).

We now can explain the origin of the peculiar name "congruent number." Fibonacci, in his book *Liber Quadratorum* (Book of Squares) from 1225, called an integer n a **congruum** if there is an integer x such that $x^2 \pm n$ are both squares. This means $x^2 - n, x^2, x^2 + n$ is a 3-term arithmetic progression of squares. Fibonacci's motivation for writing his book was the study of 3-term arithmetic progressions of integral (rather than rational) squares. Both words congruum and congruence come from the Latin *congruere*, which means "to meet together" (to congregate!). A congruum is a number related to three integer squares in a kind of agreement (having a common difference). Considering a congruum multiplied by rational squares (*e.g.*, $24 \cdot (1/2)^2 = 6$) gives the congruent numbers.

6.4 The Curve $y^2 = x^3 - n^2x$

Whether or not n is congruent is related to solvability of *pairs* of equations: first, by definition we need to solve $a^2 + b^2 = c^2$ and $(1/2)ab = n$ in positive rational numbers a, b, and c. In Section 6.3, we saw this is equivalent to solving a second pair of equations in positive rational numbers: $s^2 - r^2 = n$ and $t^2 - s^2 = n$. It turns out that the congruent number property is also equivalent to (nontrivial) rational solvability of the single equation $y^2 = x^3 - n^2x$.

This equation has three obvious rational solutions: $(0, 0)$, $(n, 0)$, and $(-n, 0)$. These are the solutions with $y = 0$.

Theorem 7. *For $n > 0$, there is a one-to-one correspondence between the following two sets:*

$$\{(a, b, c) : a^2 + b^2 = c^2, (1/2)ab = n\}, \quad \{(x, y) : y^2 = x^3 - n^2x, y \neq 0\}.$$

Mutually inverse correspondences between these sets are

$$(a, b, c) \mapsto \left(\frac{nb}{c-a}, \frac{2n^2}{c-a}\right), \quad (x, y) \mapsto \left(\frac{x^2 - n^2}{y}, \frac{2nx}{y}, \frac{x^2 + n^2}{y}\right).$$

Proof. This is a direct calculation left to the reader. We divide by $c - a$ in the first formula, and $c \neq a$ automatically since if $c = a$ then $b = 0$, but $(1/2)ab = n$ is nonzero. Restricting y to a nonzero value is necessary since we divide by y in the second formula. □

Remark. It is of course natural to wonder how the correspondence in Theorem 7 could be discovered in the first place. See the appendix.

The correspondence in Theorem 7 preserves positivity: if a, b, and c are positive then $(c - a)(c + a) = b^2 > 0$, so $c - a$ is positive and thus $x = nb/(c - a) > 0$ and $y = 2n^2/(c - a) > 0$. In the other direction, if x and y are positive then from $y^2 = x^3 - n^2x = x(x^2 - n^2)$ we see $x^2 - n^2$ has to be positive, so a, b, and c are all positive. Also, for rational $n > 0$, (a, b, c) is rational if and only if (x, y) is rational. Any solution to $a^2 + b^2 = c^2$ and $(1/2)ab = n$ needs a and b to have the same sign (since $ab = 2n > 0$), and by a sign adjustment there is a rational solution with a, b, and c all positive if there is any rational solution at all. Therefore a rational number $n > 0$ is congruent if and only if the equation $y^2 = x^3 - n^2x$ has a rational solution (x, y) with $y \neq 0$; we don't have to pay attention to whether or not x and y are positive.

A positive rational number n is *not* congruent if and only if the only rational solutions to $y^2 = x^3 - n^2x$ have $y = 0$: $(0, 0)$, $(n, 0)$, and $(-n, 0)$. For example, since 1 is not congruent (Theorem 2), the only rational solutions to $y^2 = x^3 - x$ have $y = 0$.

Example 8. Since 6 is the area of a $(3, 4, 5)$ right triangle, the equation $y^2 = x^3 - 36x$ has a rational solution with $y \neq 0$. The solution corresponding to the $(3, 4, 5)$ right triangle by Theorem 7 is $(x, y) = (12, 36)$. See Figure 6.2.

Example 9. From the rational right triangle $(3/2, 20/3, 41/6)$ with area 5, Theorem 7 gives us a rational solution to $y^2 = x^3 - 25x$: $(x, y) = (25/4, 75/8)$. If we allow sign changes on the coordinates of $(3/2, 20/3, 41/6)$, Theorem 7 will give us new rational solutions to $y^2 = x^3 - 25x$. Using the triples of the form $(\pm 3/2, \pm 20/3, \pm 41/6)$ where the first two coordinates have the same sign, the new solutions to the equation $y^2 = x^3 - 25x$ are collected in Table 6.2 and they are plotted on $y^2 = x^3 - 25x$ in Figure 6.3.

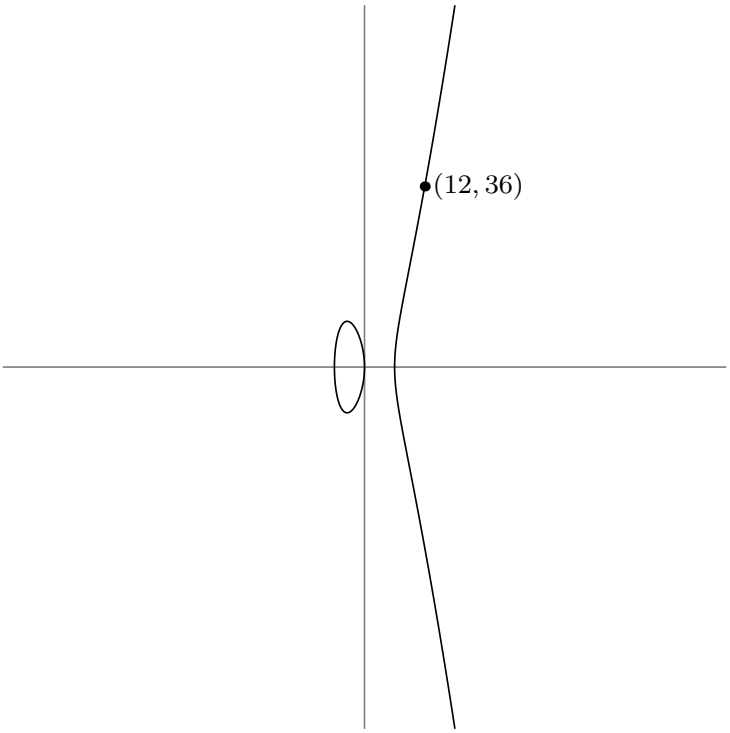


Figure 6.2: The rational point $(12, 36)$ on $y^2 = x^3 - 36x$.

Signs on $(3/2, 20/3, 41/6)$	(x, y)
$(+,+,+)$	$(25/4, 75/8)$
$(+,+,-)$	$(-4, -6)$
$(-,-,+)$	$(-4, 6)$
$(-,-,-)$	$(25/4, -75/8)$

Table 6.2: Solutions to $y^2 = x^3 - 25x$.

Example 10. A rational solution to $y^2 = x^3 - 49x$ is $(25, 120)$. Theorem 7 produces from this solution the rational right triangle $(24/5, 35/12, 337/60)$ with area 7, which we met already in Figure 6.1.

Example 11. In Table 6.1 we found two rational right triangles with area 210: $(35, 12, 37)$ and $(21, 20, 29)$. Using Theorem 7, these triangles lead to two rational solutions to $y^2 = x^3 - 210^2x$: $(1260, 44100)$ and $(525, 11025)$, respectively. In Figure 6.4, the line through $(1260, 44100)$ and $(525, 11025)$ meets the curve $y^2 = x^3 - 210^2x$ in a third point $(240, -1800)$. Its second coordinate is negative, but the point $(240, 1800)$ is also on that curve, and it leads by Theorem 7 to the new rational right triangle $(15/2, 56, 113/2)$ with area 210.

Example 12. Suppose (a, b, c) satisfies $a^2 + b^2 = c^2$ and $(1/2)ab = n$. Such a solution gives rise to seven additional ones: $(-a, -b, -c)$ and

$$(a, b, -c),\ \ (-a, -b, c),\ \ (b, a, c),\ \ (b, a, -c),\ \ (-b, -a, c),\ \ (-b, -a, -c).$$

These algebraic modifications have a geometric interpretation in terms of constructing new points from old ones on the curve $y^2 = x^3 - n^2x$ using secant lines. Say (a, b, c) corresponds to (x, y) by Theorem 7, so $y \neq 0$. From the point (x, y) on the curve, we can automatically generate a second point: $(x, -y)$. This corresponds by Theorem 7 to $(-a, -b, -c)$. What points on the curve correspond to the six remaining algebraic modifications above?

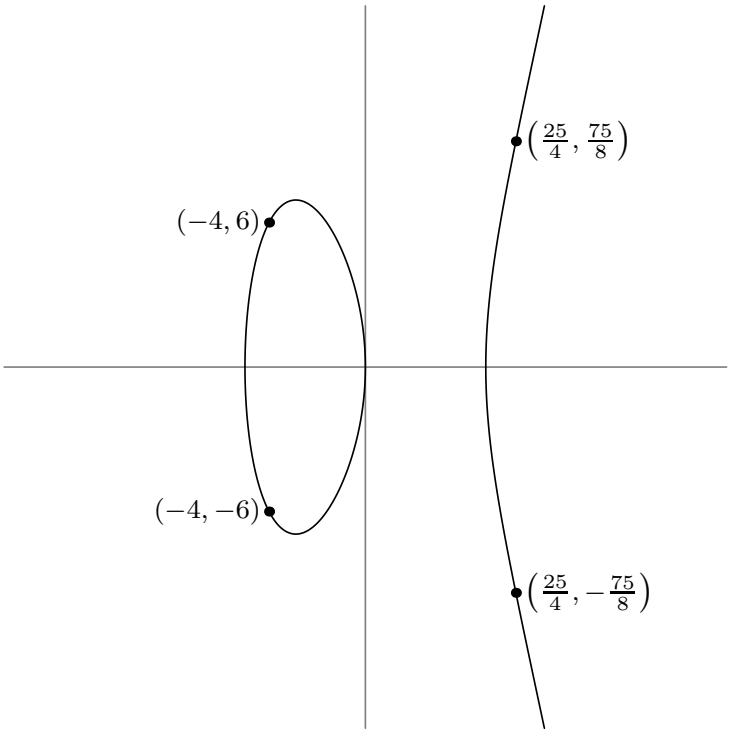


Figure 6.3: Some rational points on $y^2 = x^3 - 25x$.

Well, there are three obvious points on the curve which have nothing to do with our particular (x, y), namely $(0, 0)$, $(n, 0)$, and $(-n, 0)$. The line through (x, y) and $(0, 0)$ meets the curve in the point $(-n^2/x, -n^2y/x^2)$, which corresponds by Theorem 7 to $(a, b, -c)$. More generally, the three lines through (x, y) and each of $(0, 0)$, $(n, 0)$, and $(-n, 0)$ meet the curve in three additional points, and their reflections across the x-axis are an additional three points (which are where the lines through $(x, -y)$ and each of $(0, 0)$, $(n, 0)$, and $(-n, 0)$ meet the curve). See Table 6.3 and Figure 6.5. The corresponding triples from Theorem 7 are collected in Table 6.4 and are exactly what we were looking for.

First Point	Second Point	Third Point
(x, y)	$(0, 0)$	$(-n^2/x, -n^2y/x^2)$
$(x, -y)$	$(0, 0)$	$(-n^2/x, n^2y/x^2)$
(x, y)	$(n, 0)$	$(n(x+n)/(x-n), 2n^2y/(x-n)^2)$
$(x, -y)$	$(n, 0)$	$(n(x+n)/(x-n), -2n^2y/(x-n)^2)$
(x, y)	$(-n, 0)$	$(-n(x-n)/(x+n), 2n^2y/(x+n)^2)$
$(x, -y)$	$(-n, 0)$	$(-n(x-n)/(x+n), -2n^2y/(x+n)^2)$

Table 6.3: Third Intersection Point of a Line with $y^2 = x^3 - n^2x$.

We have seen that the following properties of a positive rational number n are equivalent:

- there is a rational right triangle with area n,
- there is a 3-term arithmetic progression of rational squares with common difference n,
- there is a rational solution to $y^2 = x^3 - n^2x$ with $y \neq 0$.

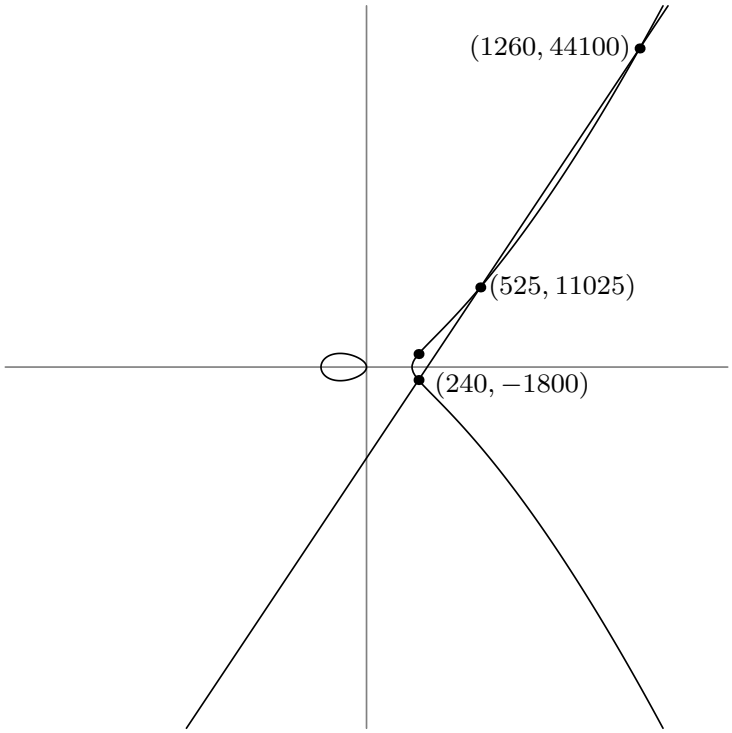


Figure 6.4: New rational point on $y^2 = x^3 - 210^2x$ from a secant line. Not drawn to scale.

Pair	Triple
(x, y)	(a, b, c)
$(x, -y)$	$(-a, -b, -c)$
$(-n^2/x, -n^2y/x^2)$	$(a, b, -c)$
$(-n^2/x, n^2y/x^2)$	$(-a, -b, c)$
$(n(x+n)/(x-n), 2n^2y/(x-n)^2)$	(b, a, c)
$(n(x+n)/(x-n), -2n^2y/(x-n)^2)$	$(-b, -a, -c)$
$(-n(x-n)/(x+n), 2n^2y/(x+n)^2)$	$(-b, -a, c)$
$(-n(x-n)/(x+n), -2n^2y/(x+n)^2)$	$(b, a, -c)$

Table 6.4: Theorem 7 and Sign Changes.

The viewpoint of the equation $y^2 = x^3 - n^2x$ lets us use the geometry of the curve to do something striking: produce a new rational right triangle with area n from two known triangles. We saw an instance of this in Example 11. Notice there is nothing in the definition of a congruent number which suggests it is possible to produce a new rational right triangle with area n from two known ones. We can even find a new rational right triangle with area n from just one such triangle, by using a tangent line in place of a secant line. Given a rational point (x_0, y_0) on $y^2 = x^3 - n^2x$ with $y_0 \neq 0$, draw the tangent line to this curve at the point (x_0, y_0). This line will meet the curve in a second rational point, and that can be converted into a new rational right triangle with area n using the correspondence of Theorem 7 (and removing any signs on a, b, c if they turn out negative.)

Example 13. In Example 11, we found a third rational right triangle from two known ones by intersecting the line through the points $(1260, 44100)$ and $(525, 11025)$ with $y^2 = x^3 - 210^2x$. We can find a new rational right triangle with area 210 from the single point $(1260, 44100)$ by

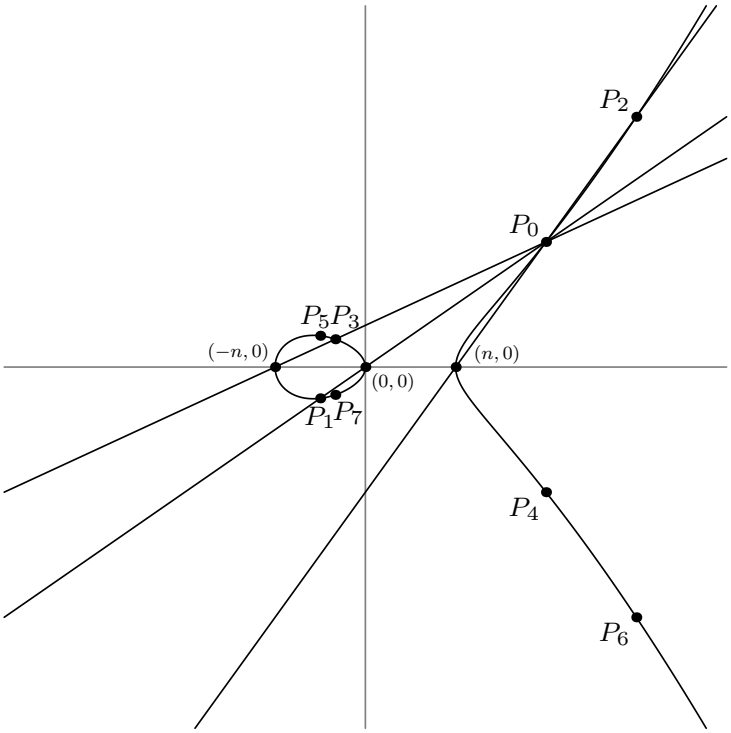


Figure 6.5: Intersecting $y^2 = x^3 - n^2x$ with lines through P_0 and $(0,0)$, $(n,0)$, $(-n,0)$, and reflected points.

using the tangent line to $y^2 = x^3 - 210^2x$ at $(1260, 44100)$. The tangent is

$$y = \frac{107}{2}x - 23310$$

and it meets the curve in the second point $(1369/4, -39997/8)$. See Figure 6.6. By Theorem 7, this point corresponds to $(a, b, c) = (-1081/74, -31080/1081, -2579761/79994)$, which after removing signs is the rational right triangle $(1081/74, 31080/1081, 2579761/79994)$, whose area is 210.

Example 14. The $(3, 4, 5)$ right triangle with area 6 corresponds to the point $(12, 36)$ on the curve $y^2 = x^3 - 36x$, as we saw already in Example 8. The tangent line to this curve at the point $(12, 36)$ is $y = (11/2)x - 30$, which meets the curve in the second point $(25/4, 35/8) = (6.25, 4.375)$. Let's repeat the tangent process on this new point. The tangent line to the curve at $(25/4, 35/8)$ has equation

$$y = \frac{1299}{140}x - \frac{6005}{112},$$

which meets the curve in the new point

$$\left(\frac{1442401}{19600}, \frac{1726556399}{2744000}\right) \approx (73.59, 629.21). \tag{6.4}$$

This is illustrated in Figure 6.7, where the second tangent line meets the curve outside the range of the picture.[1] A larger view, showing where the second tangent line meets the curve, is in Figure 6.8. (The axes in Figures 6.7 and 6.8 are not given equal scales, which is why the same tangent line in the two figures appears to have different slopes.) Using Theorem 7, $(25/4, 35/8)$ corresponds

[1]The inflection points on the curve in Figure 6.7, for $x > 0$, occur where $x = \sqrt{12(3 + 2\sqrt{3})} \approx 8.8$.

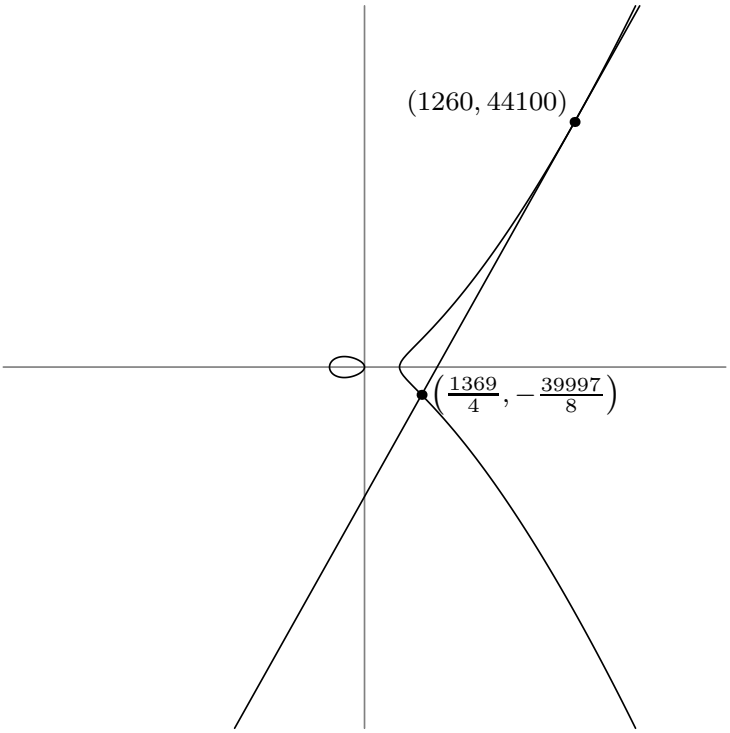


Figure 6.6: New rational point on $y^2 = x^3 - 210^2x$ from a tangent line. Not drawn to scale.

to the rational right triangle with area 6 having sides $(7/10, 120/7, 1201/70)$. The rational right triangle with area 6 corresponding to the point in (6.4) has sides

$$\left(\frac{1437599}{168140}, \frac{2017680}{1437599}, \frac{2094350404801}{241717895860}\right). \tag{6.5}$$

Armed with 3 rational right triangles with area 6, we can find 3 arithmetic progressions of rational squares using Theorem 3. The $(3, 4, 5)$ triangle, as we saw in Example 4, yields the arithmetic progression $1/4$, $25/4$, $49/4$. The $(7/10, 120/7, 1201/70)$ right triangle yields the arithmetic progression

$$\left(\frac{1151}{140}\right)^2, \ \left(\frac{1201}{140}\right)^2, \ \left(\frac{1249}{140}\right)^2.$$

The right triangle with sides in (6.5) yields the arithmetic progression

$$\left(\frac{1727438169601}{483435791720}\right)^2, \ \left(\frac{2094350404801}{483435791720}\right)^2, \ \left(\frac{77611083871}{483435791720}\right)^2.$$

All of these arithmetic progressions of squares have common difference 6.

Remark. The secant method is a way to "add" points and the tangent method is essentially the special case of "doubling" a point. These tangent and secant constructions can be used to give the rational points on $y^2 = x^3 - n^2x$ the structure of an abelian group in which, for rational $n > 0$, any rational point (x, y) with $y \neq 0$ has infinite order. (This is not at all obvious.) Therefore the curve $y^2 = x^3 - n^2x$ has infinitely many rational points as soon as it has just one rational point with $y \neq 0$, so there are infinitely many rational right triangles with area n provided there is one example and there are infinitely many 3-term arithmetic progressions of rational squares with common difference n provided there is one example. In terms of Table 6.1, this means any area

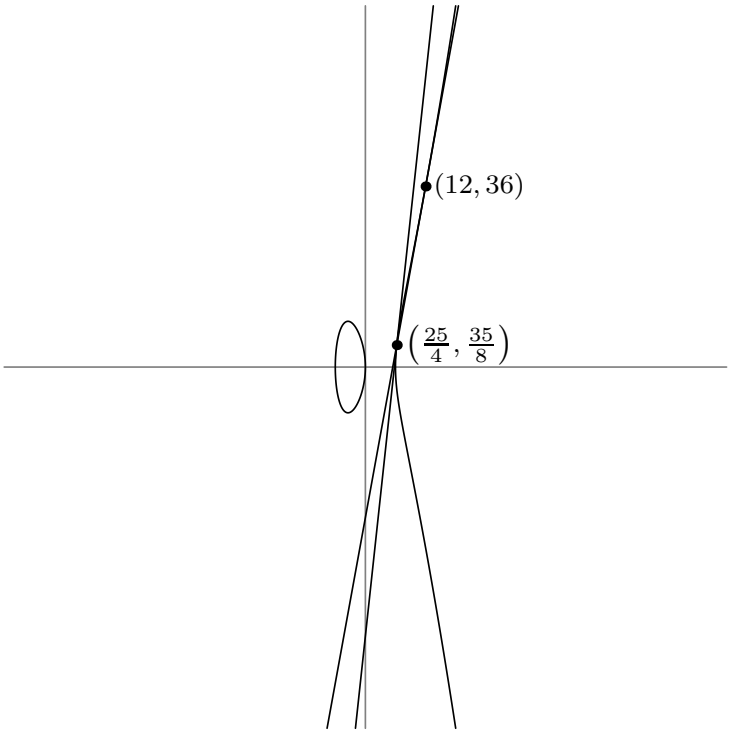


Figure 6.7: Close view of successive tangents to $y^2 = x^3 - 36x$ starting from $(12, 36)$.

arising in the table at least once will arise in the table infinitely often.[2]

The importance of thinking about congruent numbers in terms of the curves $y^2 = x^3 - n^2x$ goes far beyond this interesting construction of new rational right triangles with area n from old ones: this viewpoint in fact leads to a tentative solution of the whole congruent number problem! In 1983, Tunnell [Tu] used arithmetic properties of $y^2 = x^3 - n^2x$ (which is a particular example of an elliptic curve) to discover a previously unknown elementary necessary condition on congruent numbers and he was able to prove the condition is sufficient if a certain other conjecture is true.

Theorem 15 (Tunnell)**.** *Let n be a squarefree positive integer. Set*

$$\begin{aligned} f(n) &= \#\{(x,y,z) \in \mathbb{Z}^3 : x^2 + 2y^2 + 8z^2 = n\}, \\ g(n) &= \#\{(x,y,z) \in \mathbb{Z}^3 : x^2 + 2y^2 + 32z^2 = n\}, \\ h(n) &= \#\{(x,y,z) \in \mathbb{Z}^3 : x^2 + 4y^2 + 8z^2 = n/2\}, \\ k(n) &= \#\{(x,y,z) \in \mathbb{Z}^3 : x^2 + 4y^2 + 32z^2 = n/2\}. \end{aligned}$$

For odd n, if n is congruent then $f(n) = 2g(n)$. For even n, if n is congruent then $h(n) = 2k(n)$. Moreover, if the weak Birch and Swinnerton–Dyer conjecture is true for the curve $y^2 = x^3 - n^2x$ then the converse of both implications is true: $f(n) = 2g(n)$ implies n is congruent when n is odd and $h(n) = 2k(n)$ implies n is congruent when n is even.

The weak Birch and Swinnerton–Dyer conjecture, which we won't describe here, is one of the most important conjectures in mathematics. (It is on the list of Clay Millennium Prize problems.) Several years before Tunnell proved his theorem, Stephens [St] showed the weak Birch and Swinnerton–Dyer conjecture implies any positive integer $n \equiv 5, 6, 7 \bmod 8$ is a congruent number. Tunnell's achievement was discovering the enumerative criterion for congruent numbers and

[2]The two rational points on $y^2 = x^3 - 210^2x$ which correspond to the repetition of 210 in Table 6.1 are independent in the group law: they do not have a common multiple.

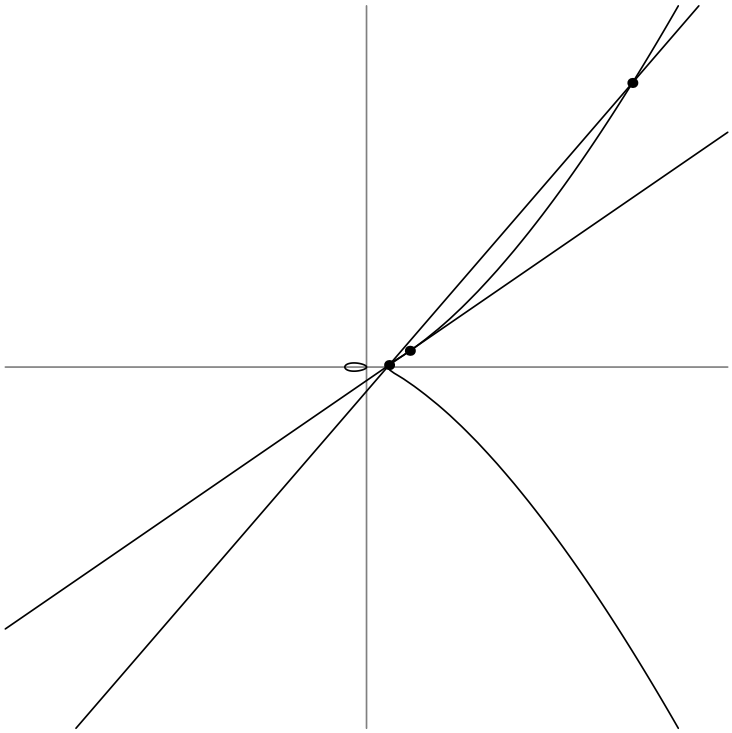


Figure 6.8: Far view of successive tangents to $y^2 = x^3 - 36x$ starting from $(12, 36)$.

its relation to the weak Birch and Swinnerton–Dyer conjecture. For background on the ideas in Tunnell's theorem, see [He] and [Ko]. In [Kn, pp. 112–114], the particular case of prime congruent numbers is considered.

Tunnell's theorem provides an unconditional method of proving a squarefree positive integer n is not congruent (show $f(n) \neq 2g(n)$ or $h(n) \neq 2k(n)$, depending on the parity of n), and a conditional method of proving n is congruent (conditional, that is, on the weak Birch and Swinnerton-Dyer conjecture for the curve $y^2 = x^3 - n^2x$).

Example 16. Since $f(1) = g(1) = 2$ and $f(3) = g(3) = 4$, we have $f(n) \neq 2g(n)$ for $n = 1$ and 3, so Tunnell's criterion shows 1 and 3 are not congruent.

Example 17. Since $h(2) = k(2) = 2$, we have $h(2) \neq 2k(2)$, so Tunnell's criterion shows 2 is not congruent.

Example 18. Since $f(5) = g(5) = 0$ and $f(7) = g(7) = 0$, we have $f(n) = 2g(n)$ for $n = 5$ and 7. Tunnell's theorem says 5 and 7 are congruent if the weak Birch and Swinnerton-Dyer conjecture is true for $y^2 = x^3 - 25x$ and $y^2 = x^3 - 49x$. Unconditionally, we saw earlier that 5 and 7 are congruent.

Example 19. Since $h(10) = 4$ and $k(10) = 4$, $h(10) \neq 2k(10)$, so Tunnell's theorem says 10 is not a congruent number.

Example 20. We will show (conditionally) that any positive integer n satisfying $n \equiv 5, 6, 7 \bmod 8$ is a congruent number. Writing $n = a^2b$ with b squarefree, a has to be odd so $n \equiv b \bmod 8$. Thus we may suppose n is squarefree. Tunnell's theorem tells us to check that $f(n) = 2g(n)$ when $n \equiv 5, 7 \bmod 8$ and $h(n) = 2k(n)$ when $n \equiv 6 \bmod 8$. Since $x^2 + 2y^2 \not\equiv 5, 7 \bmod 8$ for any integers x and y, $f(n) = 0$ and $g(n) = 0$ when $n \equiv 5, 7 \bmod 8$, so $f(n) = 2g(n)$. When $n \equiv 6 \bmod 8$ we have $n/2 \equiv 3 \bmod 4$, so $x^2 \not\equiv n/2 \bmod 4$ for any integer x. Therefore $h(n) = 0$ and $k(n) = 0$ when $n \equiv 6 \bmod 8$, so $h(n) = 2k(n)$. This shows n is congruent if the weak Birch and Swinnerton-Dyer conjecture is true for $y^2 = x^3 - n^2x$.

6.5 Acknowledgments

I thank Lucas David-Roesler and The HCMR staff for their help generating the pictures.

Appendices

6.A Discovering Theorem 7

Fix a real number $n \neq 0$. The real solutions (a, b, c) to each of the equations

$$a^2 + b^2 = c^2, \quad \frac{1}{2}ab = n, \tag{6.6}$$

describe a surface in $\mathbb{R}^3$, so it is reasonable to expect these two surfaces intersect in a curve. We want an equation for that curve, which will be $y^2 = x^3 - n^2x$ in the right choice of coordinates. Two approaches will be described, one algebraic and the other geometric. The sign on n will be irrelevant, so we allow any $n \neq 0$ rather than $n > 0$.

The algebra is simplified by introducing a cross-term in the equation $a^2 + b^2 = c^2$. Let $c = t + a$, which turns this equation into $b^2 = t^2 + 2at$, or equivalently

$$2at = b^2 - t^2. \tag{6.7}$$

Since $ab = 2n$ is nonzero, neither a nor b is 0, so we can write $a = 2n/b$ and substitute it into (6.7):

$$\frac{4nt}{b} = b^2 - t^2.$$

Multiplying through by b makes this

$$4nt = b^3 - t^2b.$$

Divide by t^3 ($t \neq 0$, as otherwise $a = c$ and then $b = 0$, but $ab = 2n \neq 0$):

$$\frac{4n}{t^2} = \left(\frac{b}{t}\right)^3 - \frac{b}{t}.$$

Multiply through by n^3:

$$\left(\frac{2n^2}{t}\right)^2 = \left(\frac{nb}{t}\right)^3 - n^2\left(\frac{nb}{t}\right).$$

Set $x = nb/t$ and $y = 2n^2/t$, so $y^2 = x^3 - n^2x$. Then $x = nb/(c - a)$ and $y = 2n^2/(c - a)$, as in Theorem 7.

We now turn to a geometric explanation of Theorem 7, taking greater advantage of the interpretation of the two equations in (6.6) as surfaces which meet in a curve. Rather than working with the equations as surfaces in $\mathbb{R}^3$, we will work in the projective space $\mathbb{P}^3(\mathbb{R})$ by homogenizing the two equations. This doesn't change the first equation in (6.6), but makes the second one $(1/2)ab = nd^2$.

Letting $[a, b, c, d]$ be the homogeneous coordinates of a typical point in $\mathbb{P}^3(\mathbb{R})$, the two equations

$$a^2 + b^2 = c^2, \quad \frac{1}{2}ab = nd^2 \tag{6.8}$$

each define surfaces in $\mathbb{P}^3(\mathbb{R})$. Let C be the intersection of these surfaces (a curve). There are points on C with $b = 0$, namely $[a, b, c, d] = [1, 0, \pm 1, 0]$. These points are not in the usual affine space inside $\mathbb{P}^3(\mathbb{R})$, and we will use one of these points in a geometric construction.

Let's project through the point $P := [1, 0, 1, 0]$ to map C to the plane

$$\Pi := \{[0, b, c, d]\}$$

and find the equation for the image of C in this plane. The point P lies on C and not in Π. For each $Q \in C$ other than P, the line $\overline{PQ}$ in $\mathbb{P}^3(\mathbb{R})$ meets Π in a unique point. Call this point $f(Q)$. When $Q = P$, intersect the tangent line to C at P with the plane Π to define $f(P)$. We have defined a function $f\colon C \to \Pi$.

Computing a formula for f necessitates a certain amount of computation to see what happens. Suppose first that $Q = [a, b, c, d]$ is not P. The line through P and Q is the set of points

$$[\lambda + \mu a, \mu b, \lambda + \mu c, \mu d],$$

which meets Π where $\lambda = -\mu a$, making

$$f(Q) = [0, \mu b, \mu(c - a), \mu d] = [0, b, c - a, d].$$

As for $f(P)$, the tangent planes to each of the surfaces $a^2 + b^2 = c^2$ and $(1/2)ab = nd^2$ in $\mathbb{P}^3(\mathbb{R})$ at the point P are the planes $a = c$ and $b = 0$, so the tangent line at P is the set of points

$$[a, 0, a, d],$$

which meets Π in $[0, 0, 0, 1]$, so $f(P) = [0, 0, 0, 1]$. Thus

$$f([a, b, c, d]) = \begin{cases} [0, b, c - a, d], & \text{if } [a, b, c, d] \neq [1, 0, 1, 0], \\ [0, 0, 0, 1], & \text{if } [a, b, c, d] = [1, 0, 1, 0]. \end{cases}$$

As an exercise, check f is injective. (Hint: Since $(1/2)ab = nd^2$, b and d determine a if $b \neq 0$.)

All points in the plane Π have first coordinate 0. Identify Π with $\mathbb{P}^2(\mathbb{R})$ by dropping this coordinate, which turns f into the function $g\colon C \to \mathbb{P}^2(\mathbb{R})$ where

$$g([a, b, c, d]) = \begin{cases} [b, a - c, d], & \text{if } [a, b, c, d] \neq [1, 0, 1, 0], \\ [0, 0, 1], & \text{if } [a, b, c, d] = [1, 0, 1, 0]. \end{cases} \tag{6.9}$$

See Figure 6.9, where P is located "at infinity" in a vertical direction.

We have mapped our curve C to the projective plane $\mathbb{P}^2(\mathbb{R})$. What is an equation for the image $g(C)$? For $Q = [a, b, c, d]$ on C, write $g(Q) = [x, z, y]$. (This ordering of the coordinates will make formulas come out close to the expected way more quickly.) When $Q \neq [1, 0, 1, 0]$ (that is, $a \neq c$), (6.9) says we can use $x = b$, $y = d$, and $z = c - a \neq 0$.[3] The equations in (6.8) become $a^2 + x^2 = (a + z)^2$ and $(1/2)ax = ny^2$, so

$$x^2 = 2az + z^2, \quad ax = 2ny^2.$$

Since $z \neq 0$, we can solve for a in the first equation, so a is determined by x, y, and z. Multiplying the first equation by x and the second by $2z$, $x^3 = 2axz + xz^2 = 4ny^2z + xz^2$. Thus

$$4ny^2z = x^3 - xz^2.$$

Set $X = x$, $Y = 2ny$, and $Z = z/n$ to find $Y^2Z = X^3 - n^2XZ^2$, which is the homogeneous form of $Y^2 = X^3 - n^2X$.

Tracing this correspondence out explicitly from the start, if we begin with $[a, b, c, d]$ on C where $d \neq 0$ (the standard affine part of C), its image $[X, Z, Y]$ in $\mathbb{P}^2(\mathbb{R})$ is

$$\left[b, \frac{c - a}{n}, 2nd\right] = [nb, c - a, 2n^2d] = \left[\frac{nb}{c - a}, 1, \frac{2n^2d}{c - a}\right].$$

[3]The cross term $t = c - a$ in the algebraic method is precisely z, so now we get a geometric interpretation of this cross term as a coordinate in a projection map to a plane.

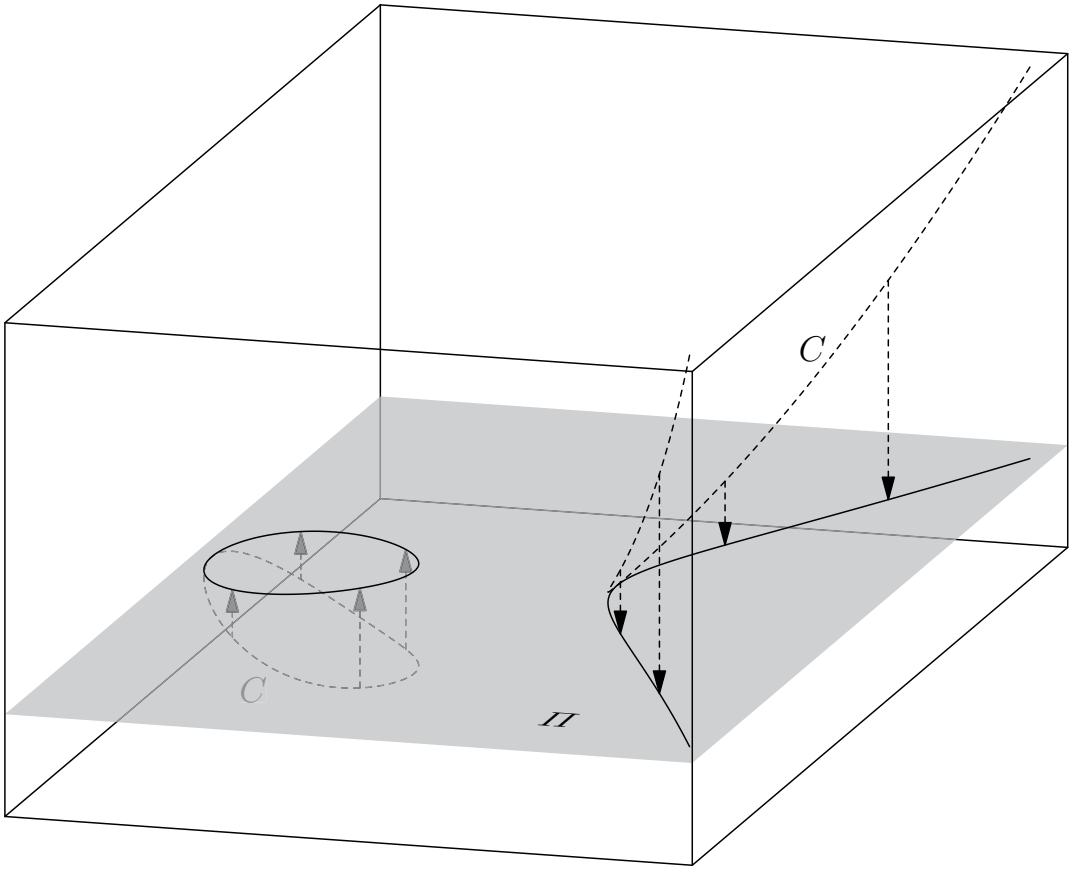


Figure 6.9: Projection in (6.9) through point P at infinity from curve C to $\Pi \cong \mathbb{P}^2(\mathbb{R})$.

Since $d \neq 0$ implies $a \neq c$, using inhomogeneous coordinates with middle coordinate 1 in $\mathbb{P}^2(\mathbb{R})$ the point (a, b, c) goes to $(nb/(c-a), 2n^2/(c-a))$, which is the transformation in Theorem 7.

As an exercise in these techniques, consider the problem of classifying triangles with a given area $n > 0$ and a given angle θ. (Taking $\theta = \pi/2$ is the congruent number problem.) Let a, b, c be the side lengths of the triangle, with c the length of the edge opposite the angle θ. The equations in (6.6) are replaced by

$$a^2 + b^2 - 2ab\cos\theta = c^2, \quad \frac{1}{2}ab\sin\theta = n. \tag{6.10}$$

(If there is a solution with rational a, b, c, and n then $\cos\theta$ and $\sin\theta$ must be rational.) Show the solutions (a, b, c) of (6.10) are in one-to-one correspondence with the solutions (x, y) of the equation

$$y^2 = x^3 + \frac{2n\cos\theta}{\sin\theta}x^2 - n^2x = x\left(x + n\frac{\cos\theta + 1}{\sin\theta}\right)\left(x + n\frac{\cos\theta - 1}{\sin\theta}\right),$$

with $y \neq 0$. The correspondence should specialize to that in Theorem 7 when $\theta = \pi/2$.

6.B Other Diophantine Equations

In Table 6.5, the first two columns show how to convert the sides (a, b, c) of a rational right triangle with area 1 into a positive rational solution of the equation $y^2 = x^4 - 1$ and conversely. (These correspondences are *not* inverses, but they do show a positive rational solution in the first column leads to a positive rational solution in the second column, and conversely.) The last two columns give a (bijective) correspondence between rational right triangles with area 2 and positive rational

solutions of $y^2 = x^4 + 1$. So showing 1 and 2 are not congruent numbers is the same as showing the equations $y^2 = x^4 \pm 1$ don't have solutions in positive rational numbers.

$a^2+b^2=c^2$, $\frac{1}{2}ab=1$	$y^2 = x^4 - 1$	$a^2+b^2=c^2$, $\frac{1}{2}ab=2$	$y^2 = x^4 + 1$
$x = c/2$	$a = y/x$	$x = a/2$	$a = 2x$
$y = \lvert a^2 - b^2\rvert/4$	$b = 2x/y$	$y = ac/4$	$b = 2/x$
	$c = (x^4+1)/xy$		$c = 2y/x$

Table 6.5: Correspondences between rational right triangles with area 1 or 2 and $y^2 = x^4 \pm 1$.

A positive rational solution (x, y) to $y^2 = x^4 \pm 1$ can be turned into a positive integral solution (u, v, w) of $w^2 = u^4 \pm v^4$ by clearing a common denominator, and we can go in reverse by dividing by v^4. That 1 and 2 are not congruent is therefore the same as the equations $w^2 = u^4 \pm v^4$ having no positive integer solutions. The reader is referred to [Bu, pp. 252–256] for a proof by descent that $w^2 = u^4 \pm v^4$ has no positive integer solutions.

That the congruent number property for 1 and 2 is equivalent to the solvability of a single equation in positive rational numbers ($y^2 = x^4 - 1$ for 1 and $y^2 = x^4 + 1$ for 2) generalizes: n is congruent if and only if $y^2 = x^4 - n^2$ has a positive rational solution and if and only if $y^2 = x^4 + 4n^2$ has a positive rational solution. See Table 6.6, where the first two columns turn rational right triangles with area n into positive rational solutions of $y^2 = x^4 - n^2$ and conversely, and the last two columns do the same with $y^2 = x^4 + 4n^2$. As in Table 6.5, the correspondences in the first two columns of Table 6.6 are not inverses of each other, but the correspondences in the last two columns are inverses. (When $n = 2$ the equation in Table 6.6 is $y^2 = x^4 + 16$ rather than $y^2 = x^4 + 1$ as in Table 6.5. We can easily pass from the former to the latter by replacing y with $4y$ and x with $2x$.) The equivalence of n being congruent with $y^2 = x^4 - n^2$ having a positive rational solution is due to Lucas (1877).

$a^2+b^2=c^2$, $\frac{1}{2}ab=n$	$y^2 = x^4 - n^2$	$a^2+b^2=c^2$, $\frac{1}{2}ab=n$	$y^2 = x^4 + 4n^2$
$x = c/2$	$a = y/x$	$x = a$	$a = x$
$y = \lvert a^2 - b^2\rvert/4$	$b = 2nx/y$	$y = ac$	$b = 2n/x$
	$c = (x^4+n^2)/xy$		$c = y/x$

Table 6.6: More correspondences between rational right triangles and Diophantine equations.

We pulled the equations $y^2 = x^4 - n^2$ and $y^2 = x^4 + 4n^2$ out of nowhere. How could they be discovered? The arithmetic progression viewpoint on congruent numbers (Theorem 3) leads to one of them. If n is congruent, there are rational squares r^2, s^2, and t^2 with $s^2 - r^2 = n$ and $t^2 - s^2 = n$. Then $r^2 = s^2 - n$ and $t^2 = s^2 + n$, so multiplication gives $(rt)^2 = s^4 - n^2$ and we've solved $y^2 = x^4 - n^2$ in positive rational numbers.

Remark. For $t \neq 0$, solutions to $y^2 = x^4 + t$ and to $Y^2 = X^3 - 4tX$ are in a one-to-one correspondence, by $(x, y) \mapsto (2t/(y - x^2), 4tx/(y - x^2))$ and $(X, Y) \mapsto (Y/2X, (Y^2 + 8tX)/4X^2)$. In particular, solutions to $y^2 = x^4 - n^2$ correspond to solutions to $Y^2 = X^3 + (2n)^2X$, which is *not* the equation $Y^2 = X^3 - (2n)^2X$ and thus isn't related to whether or not $2n$ is a congruent number. Explicit examples show the lack of a general connection between n and $2n$ being congruent: 5 is congruent but 10 is not, while 3 is not congruent but 6 is.

References

[Bu] D. M. Burton: *Elementary Number Theory*, 6th ed. New York: McGraw-Hill 2007.

[Co] W. A. Coppel: *Number Theory: An Introduction to Mathematics. Part B*. New York: Springer-Verlag 2006.

[Di] L. E. Dickson: *History of the Theory of Numbers*, Vol. II. New York: Chelsea 1952.

[He] G. Henniart: Congruent Numbers, Elliptic Curves, and Modular Forms, transl. F. Lemmermeyer at `http://www.fen.bilkent.edu.tr/~franz/publ.html`.

[Kn] A. Knapp: *Elliptic Curves*. Princeton: Princeton Univ. Press 1992.

[Ko] N. Koblitz: *Introduction to Elliptic Curves and Modular Forms*, 2nd ed. New York: Springer-Verlag 1993.

[St] N. M. Stephens: Congruence properties of congruent numbers, *Bull. London Math. Soc.* **7** (1975), 182–184.

[Tu] J. Tunnell: A Classical Diophantine Problem and Modular Forms of Weight $3/2$, *Invent. Math.* **72** (1983), 323–334.

FEATURE

7

MATHEMATICAL MINUTIAE

Quadratic Reciprocity by Group Theory

Tim Kunisky[†]
Livingston High School '10
Livingston, NJ 07039
tkunisky@gmail.com

For p a prime, consider $(\mathbb{Z}/p\mathbb{Z})^\times$, the multiplicative group of the nonzero integers modulo p. We know that exactly half of the elements are squares, and want to find them. The Legendre symbol $\left(\frac{a}{p}\right)$ is defined to be 1 if a is a square in $(\mathbb{Z}/p\mathbb{Z})^\times$, -1 otherwise. Besides simply squaring the integers from 1 to $p-1$ to see if $\left(\frac{a}{p}\right) = 1$, we can also use Euler's Criterion, which states that $\left(\frac{a}{p}\right) = a^{\frac{p-1}{2}} \mod p$. However, the most elegant way uses the law of quadratic reciprocity, first proven by Gauss. It states that if p and q are odd primes then

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

We will prove this result using elementary group theory. Consider the group $G = (\mathbb{Z}/p\mathbb{Z})^\times \times (\mathbb{Z}/q\mathbb{Z})^\times$ for p, q odd primes. Note that $A = \{(1,1),(-1,-1)\}$ is a normal subgroup of G, and let $H = G/A$ be the quotient group. We will find two equivalent expressions for the product of all elements of H by considering coset representatives for A.

Any $(a,b) \in G$ can be written uniquely as $(a, \pm b')$ where $1 \le a \le p-1$ and $1 \le b' \le \frac{q-1}{2}$. Since negating a does not change the possibilities for the first coordinate, $S = \{(x,y) \mid 1 \le x \le p-1, 1 \le y \le \frac{q-1}{2}\}$ is a set of coset representatives for A. Taking the product of all elements of this set gives

$$\left((p-1)!^{\frac{q-1}{2}}, \left(\tfrac{q-1}{2}\right)!^{p-1}\right)$$

but we know that in $\mathbb{Z}/q\mathbb{Z}$

$$\left(\tfrac{q-1}{2}\right)!^2 = (-1)^{\frac{q-1}{2}}(q-1)!$$

and therefore

$$\left(\tfrac{q-1}{2}\right)!^{p-1} = \left(\left(\tfrac{q-1}{2}\right)!^2\right)^{\frac{p-1}{2}} = \left((-1)^{\frac{q-1}{2}}(q-1)!\right)^{\frac{p-1}{2}} = (-1)^{\frac{q-1}{2}\cdot\frac{p-1}{2}}(q-1)!^{\frac{p-1}{2}}.$$

So the product can be rewritten as

$$\left((p-1)!^{\frac{q-1}{2}}, (-1)^{\frac{q-1}{2}\cdot\frac{p-1}{2}}(q-1)!^{\frac{p-1}{2}}\right).$$

[†]Currently, Tim is in his junior year at Livingston High School in Livingston, New Jersey. Largely influenced by his time at PROMYS, his mathematical interests have been oriented towards number theory along with algebra, though he has made attempts at studying analysis independently more recently as well. His first experience with extracurricular study was a deeper exploration of calculus, but his interests have shifted significantly to more foundational branches, including abstract algebra and set theory. Tim is fascinated with elegant proofs of simple or well-known theorems—a passion that resulted in the creation of this proof. In the next few years, he hopes to narrow his interests and pursue mathematics in college and beyond.

Next, we apply the Chinese Remainder Theorem to find another set of coset representatives, namely the set

$$T = \left\{(k \bmod p, k \bmod q) \mid k = 1, 2, \cdots, \tfrac{pq-1}{2}; (k, pq) = 1\right\}.$$

Clearly under multiplication by $(-1, -1)$ the elements of $\mathbb{Z}/pq\mathbb{Z}$ over $\frac{pq-1}{2}$ are included, and generate all elements of G that are not in T. Therefore, this is a second set of coset representatives.

Denote the product of these ordered pairs by (r, s). Then, r is the product of all k taken modulo p, and s is the same product but modulo q. Since we require $(k, pq) = 1$, to calculate r we may exclude all multiples of p, then divide out all multiples of q:

$$r = \frac{\left(\prod_{i=1}^{p-1} i\right)\left(\prod_{i=1}^{p-1} p+i\right)\cdots\left(\prod_{i=1}^{p-1}\left(\frac{q-1}{2}-1\right)p+i\right)\left(\prod_{i=1}^{\frac{p-1}{2}}\frac{q-1}{2}p+i\right)}{(1\cdot q)(2\cdot q)\cdots\left(\frac{p-1}{2}\cdot q\right)}$$

By manipulating the terms and applying Euler's Criterion, we find:

$$r = \frac{(p-1)!^{\frac{q-1}{2}}\left(\frac{p-1}{2}\right)!}{q^{\frac{p-1}{2}}\left(\frac{p-1}{2}\right)!} = \frac{(p-1)!^{\frac{q-1}{2}}}{q^{\frac{p-1}{2}}} = \frac{(p-1)!^{\frac{q-1}{2}}}{\left(\frac{q}{p}\right)} = (p-1)!^{\frac{q-1}{2}}\left(\frac{q}{p}\right)$$

We also have a symmetric expression for s:

$$s = (q-1)!^{\frac{p-1}{2}}\left(\frac{p}{q}\right)$$

So by equating this with the product from the previous calculation we find:

$$\left((p-1)!^{\frac{q-1}{2}}, (-1)^{\frac{q-1}{2}\cdot\frac{p-1}{2}}(q-1)!^{\frac{p-1}{2}}\right) = \left((p-1)!^{\frac{q-1}{2}}\left(\frac{q}{p}\right), (q-1)!^{\frac{p-1}{2}}\left(\frac{p}{q}\right)\right)$$

Therefore,

$$\left(1, (-1)^{\frac{q-1}{2}\cdot\frac{p-1}{2}}\right) = \left(\left(\frac{q}{p}\right), \left(\frac{p}{q}\right)\right).$$

Since we have been working in G/A, this is only accurate up to sign. But if we multiply the components of the ordered pairs together having both negative will make no difference, so we have the desired equation:

$$\left(\frac{q}{p}\right)\left(\frac{p}{q}\right) = (-1)^{\frac{q-1}{2}\cdot\frac{p-1}{2}}$$

FEATURE

8

STATISTICS CORNER

Conformal Invariance in the Scaling Limit of Critical Planar Percolation

Nike Sun[†]
Harvard University '09
Cambridge, MA 02138
nsun@fas.harvard.edu

8.1 Introduction to percolation theory

Percolation theory was originally developed to model the flow of liquid through a disordered porous medium. The classic example comes from coffee-making: If water is poured through coffee grounds, we would like to find out what the wet portion of the grounds might look like.

We can model the material as a graph Λ, with vertex set V and edge set E. (We will be concerned only with the case where Λ is undirected.) Each vertex is a particle of coffee, and edges join vertices corresponding to adjacent particles. In the standard terminology of percolation theory, vertices and edges are referred to as **sites** and **bonds** respectively. We can then model percolation as a random binary function on the graph: In **site percolation**, each site is independently set to be **open** or **closed** (wet or dry) with probability p; the open sites induce the **open subgraph** of Λ. **Bond percolation** is defined analogously, the only modification being that we select a random subset of **open bonds** of E. A choice of open and closed sites (or bonds) is called a **(percolation) configuration**. Percolation theory is specifically concerned with the connected components of the open subgraph, or **open clusters** — the wet portions of the coffee which are now sticking together.

For further background on this subject see Grimmett [Gr] and Bollobás and Riordan [BR]. Besides giving us insights into what happens inside our coffee-makers, percolation theory has been applied to study earthquakes and fault patterns, groundwater flow in rock, reactions in evolving porous media, random electrical networks, and semiconductors [Sa].

The goal of this article is to describe a surprising connection between the scaling limit of percolation and conformal maps. This article assumes some knowledge of complex analysis (as in [Ah] or [SS]) and basic probability theory.

8.1.1 Critical percolation probabilities

Let $\mathbb{P}_p$ denote the probability measure induced by percolation at probability p on the space of subgraphs of Λ. (In simpler terms, for any event E which is determined by the states of any or all of the sites in the graph — for example, the event that there is an infinite open cluster — $\mathbb{P}_p(E)$ denotes the probability that E will occur if each site is chosen to be open at probability p.) For $x \in \Lambda$ we can consider C_x, the **open cluster at** x, which is the connected component of the open subgraph containing x. (In particular, $C_x = \varnothing$ if and only if x is closed.) We define two quantities of interest at x:

$$\theta_x(p) = \mathbb{P}_p(|C_x| = \infty) \text{ and } \chi_x(p) = \mathbb{E}_p(|C_x|),$$

where $\mathbb{E}_p$ denotes expectation with respect to $\mathbb{P}_p$. θ_x and χ_x are nondecreasing in p. If Λ is a connected graph, then for any $x, y \in \Lambda$, $\theta_x(p)$ and $\theta_y(p)$ must be both positive or both zero,

[†]Nike Sun, Harvard '09, is a mathematics concentrator living in Winthrop House. She is also enrolled in a concurrent masters program in statistics.

and $\chi_x(p)$ and $\chi_y(p)$ must be both finite or both infinite. Therefore we can define two **critical probabilities** for the graph Λ,

$$p_H = \inf\{p : \theta_x(p) > 0\}, \;\; p_T = \inf\{p : \chi_x(p) = \infty\}.$$

(The subscript of p_H refers to Hammersley while that of p_T refers to Temperley.) If $p > p_H$, at every $x \in \Lambda$ there is a positive probability that C_x is infinite; in particular, by the Kolmogorov 0-1 Law, an infinite connected component exists with probability 1 (see e.g. [Ro]).

We always have $p_T \leq p_H$ for a given percolation model; $p_T < p_H$ can occur if, for some values of p, $|C_x|$ is finite with probability 1 but has a heavy-tailed distribution. Menshikov proves that under some uniformity conditions on the structure of Λ, for $p < p_H$ the distribution of $|C_x|$ has an almost exponential tail, so that $\chi_x(p)$ must be finite; and this is enough to conclude $p_T = p_H$. For example, $p_T = p_H = 1/2$ for Λ equal to $\mathbb{Z}^2$ or T (the triangular lattice) [BR]. We will see later that the most interesting behavior occurs for models at critical probabilities — when the medium is neither under- nor over-saturated.

8.1.2 Crossing probabilities in the scaling limit

For the remainder of this discussion, we will restrict ourselves to **site percolation on a planar lattice**. Planar lattices all satisfy the conditions of Menshikov's theorem, so their two critical probabilities are equal, and we denote them both by p_c. In particular, the main result of this section, due to Smirnov, concerns the triangular lattice T, shown with its dual hexagonal lattice H in Figure 8.1. Site percolation on a lattice can be visualized as **face percolation** (the shaded hexagons) on its dual.

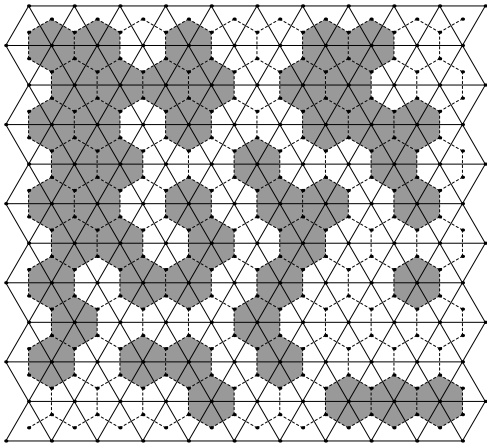


Figure 8.1: Triangular lattice T, hexagonal dual H

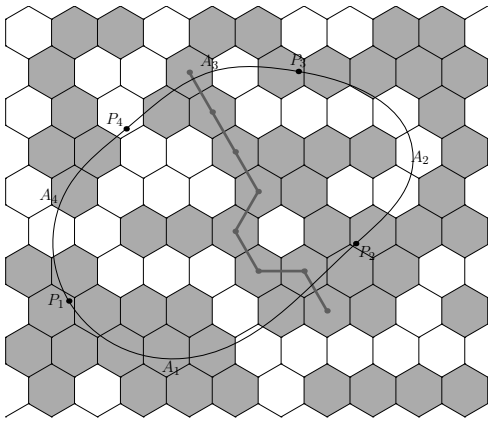


Figure 8.2: Open crossing of D (dark) in lattice δT

Let $D \subset \mathbb{C}$ be a simply connected, bounded domain, whose boundary is a Jordan curve Γ. Let z_i, $1 \leq i \leq 4$, be distinct boundary points of D, appearing in this cyclic order as Γ is traversed counterclockwise; then $D_4 = (D; z_1, z_2, z_3, z_4)$ is a **4-marked domain**. Let $A_i = A_i(D_4)$ be the arc of Γ from z_i to z_{i+1}, where the indices are taken modulo 4.

Let Λ be a planar lattice; we wish to analyze the structure of percolation on $\delta\Lambda$, the rescaled lattice, within D_4 as $\delta \to 0$. This is the notion of a **scaling limit**. By an **open crossing of D from A_1 to A_3 in $\delta\Lambda$**, we mean an open path $v_0 v_1 \cdots v_t$ in $\delta\Lambda$ such that $v_1, \cdots, v_{t-1} \in D$, $v_0, v_t \notin D$, and $v_0 v_1$ meets A_1 while $v_{t-1} v_t$ meets A_3; see Figure 8.2. Considering either site or bond percolation on Λ, for $\delta > 0$ and $0 < p < 1$, we can then define

$$P_\delta(D_4, \Lambda, p) = \mathbb{P}_p(D_4 \text{ has an open crossing in } \delta\Lambda).$$

We are then interested in $\lim_{\delta \to 0} P_\delta(D_4, \Lambda, p)$.

In fact, for almost all values of p the limit is uninteresting: Menshikov's theorem can be applied to show that for $p < p_c$ the limit is 0, and for $p > p_c$ the limit is 1. The only interesting behavior occurs when $p = p_c$; we therefore restrict ourselves to **critical percolation**, and denote $\mathbb{P}_\delta(D_4, \Lambda) = \mathbb{P}_\delta(D_4, \Lambda, p_c)$, and $\pi(D_4, \Lambda) = \lim_{\delta \to 0} P_\delta(D_4, \Lambda)$. Based on experiments, Langlands, Pouliot, and Saint-Aubin [LPSA] made the following conjecture:

Conjecture 1 ([LPSA]). *The limit $\pi(D_4, \Lambda)$ is defined, lies in $(0, 1)$, and is **conformally invariant**: If $D_4 = (D; z_1, z_2, z_3, z_4)$ and $D_4' = (D', z_1', z_2', z_3', z_4')$ are conformally equivalent 4-marked domains (there is a conformal map $\varphi : D \to D'$ which extends continuously to ∂D with $P_i \mapsto P_i'$), then $\pi(D_4, \Lambda) = \pi(D_4', \Lambda)$.*

To date this result has only been proven for the triangular lattice T, a fairly recent result due to Smirnov [Sm, Sm]. We will discuss this result, and a more general result also due to Smirnov, below. However, we first provide some motivation for why the scaling limit should be conformally invariant.

8.2 Conformal invariance of planar Brownian motion

Brownian motion is the scaling limit of **simple random walk** on $\mathbb{Z}$, the process which starts at some integer and at each step moves by ± 1 with equal probability, independently of all previous steps. Brownian motion is a **continuous stochastic process** $(B_t)_{t \geq 0}$ with normally distributed increments (e.g. by the central limit theorem), and with disjoint increments independent. The standard definition takes $B_t \sim \mathcal{N}(0, t)$, so that $B_t - B_s \sim \mathcal{N}(0, t - s)$ for $t > s$. A very readable introduction to Brownian motion can be found in Steele's book [St]; see also [RW, Va].

Now let $(B_t)_{t \geq 0}$ denote **complex Brownian motion** started at some $z \in \mathbb{C}$, that is, the real and imaginary parts B_t^1, B_t^2 of (B_t) are independent Brownian motions. We will denote the probability measure for (B_t) by $\mathbb{P}_z$, where the subscript denotes the starting point of the process.

A translation of $\mathbb{C}$ of course maps (B_t) to another complex Brownian motion. Also, since the bivariate normal distribution has rotational symmetry, a rotation of the complex plane maps (B_t) to another complex Brownian motion. In fact, we can say much more. The following observation, due to Paul Lévy, essentially tells us that complex Brownian motion behaves as nicely as possible with respect to holomorphic transformations of their domain.

Theorem 2 ([La, Le]). *Let U be a domain in $\mathbb{C}$ with $z_0 \in U$, and let B_t be a complex Brownian motion started at z_0. Set $\tau_U = \inf\{t \geq 0 \mid B_t \notin U\}$ to be the hitting time of $\mathbb{C} \setminus U$. Let $f : U \to \mathbb{C}$ be a non-constant holomorphic map, and define $Y_t = f(B_t)$ for $0 \leq t \leq \tau_U$. Then*

$$Y_{\sigma(t)}, \text{ where } \sigma^{-1}(t) = \int_0^t |f'(B_s)|^2 \, ds,$$

has the distribution of a standard Brownian motion.

Proof. We sketch the basic argument given by Lévy [Le], using the heuristic that a stochastic process is determined by its behavior locally near every point. The result is then intuitive since locally near B_t, the map f resembles rotation-dilation by $\lambda_t = f'(B_t)$. Rotations of Brownian motions are still Brownian motions, so locally Y_t looks like $|\lambda_t| B_t$. For $c \in \mathbb{R}$, if B_t is complex Brownian motion, then $cB_{t/c^2} = B_t''$ is another complex Brownian motion, by a simple calculation. Since $Y_{\sigma(t)}$ locally looks like $|\lambda_{\sigma(t)}| B_{\sigma(t)}$, $\sigma(t)$ should locally look like $t/|\lambda_{\sigma(t)}|^2$, that is, we should set $\sigma'(t) = 1/|f'(B_{\sigma(t)})|^2$. This is accomplished if $(\sigma^{-1})'(t) = |f'(B_t)|^2$. □

In words, the theorem says that **a holomorphic map of a complex Brownian motion is another complex Brownian motion**, up to a (random) time change. In particular, the curve traced out by $f(B_t)$, $0 \leq t \leq \tau_U$, is indistinguishable from a curve traced out by a complex Brownian motion. Most modern proofs of this result use Itō's lemma, together with harmonicity of the real and imaginary parts of φ [Ga, La]. Because Brownian motion occurs as the scaling limit of simple random walk, this result is one of the strongest motivations for the study of conformally invariant scaling limits coming from general discrete processes.

A further connection between Brownian motion and harmonic functions lies in the following result, due to Kakutani [Ka]. The form of the statement below is from Lawler [La], and shows that Brownian motion gives a solution to the **Dirichlet problem**.

Proposition 3 ([Ka, La])**.** *Let $D \subset \mathbb{C}$ be a bounded Jordan domain (not necessarily simply connected), and let $f : \partial D \to \mathbb{R}$ be bounded and measurable. Let $\tau_D = \inf\{t > 0 : B_t \notin D\}$ be the hitting time of $\mathbb{C} \setminus D$. Define $u : \overline{D} \to \mathbb{R}$ by*

$$u(z) = \begin{cases} f(z) & \text{if } z \in \partial D. \\ \mathbb{E}_z f(B_{\tau_D}) & \text{if } z \in D. \end{cases}$$

Then u is a bounded, harmonic function in D and is continuous at all points $z \in \partial D$ at which f is continuous.

Proof. $\|u\|_\infty \leq \|f\|_\infty < \infty$ (by assumption), so u is bounded. To show that u is harmonic, it suffices to check the mean-value property (see [Ah]),

$$u(z) = \frac{1}{2\pi} \int_0^{2\pi} u(z + re^{i\theta})\, d\theta,$$

for any $z \in D$ and any $r > 0$ such that $\overline{D_r(z)} = \{|w - z| \leq r\}$ is contained in D. Let σ denote the hitting time of $\mathbb{C} \setminus D_r(z_0)$. By the rotational symmetry of Brownian motion, B_σ has the uniform distribution on $\partial D_r(z)$ with respect to $\mathbb{P}_z$, so the above equals

$$\mathbb{E}_z(u(B_\sigma)) = \mathbb{E}_z\left[\mathbb{E}_{B_\sigma}(f(B_{\tau_D}))\right] = \mathbb{E}_z\left[\mathbb{E}_z(f(B_{\tau_D})|B_\sigma)\right] = \mathbb{E}_z f(B_{\tau_D}) = u(z),$$

where the second equality uses the strong Markov property, and the third is the law of iterated expectations. This proves that u is harmonic, and the continuity result follows from the boundedness of f together with the regularity of the domain. □

Brownian motion is the simplest non-trivial example of the scaling limit of a discrete process, and yet most known examples of stochastic processes are related to Brownian motion. Kakutani's result therefore suggests a very general connection between stochastic processes and complex analysis.

8.3 Smirnov's theorem

Smirnov [Sm, Sm] proves conformal invariance of crossing probabilities for the triangular lattice. The proof approximates D by a succession of 4-marked **discrete domains** G_δ in δT: Each $G = G_\delta$ has vertices v_i ($1 \leq i \leq 4$) marked on its internal boundary to approximate the P_i; the vertices demarcate arcs $A_i(G)$ on the internal boundary, and the external boundary can be partitioned into corresponding arcs $A_i^+(G)$.

The main idea is to express crossing probabilities for a 4-marked discrete domain G_4 in terms of **separating probabilities** for the 3-marked discrete domain $G_3 = (G; v_1, v_2, v_3)$ obtained by dropping v_4: For $z \in \delta H \cap D$ (so $z \in D$ is the center of a triangle in δT), we can define

$$s_\delta^i(z) = \mathbb{P}(\exists \text{ open } A_{i-1}(G_3)\text{–}A_i(G_3) \text{ path in } \delta\Lambda \text{ separating } z \text{ from } A_{i+1}^+(G_3)),$$

for $i = 1, 2, 3$, where the indices are taken modulo 3. That is, $s_\delta^i(z)$ is the probability there is a (simple) open path joining the two boundary arcs meeting v_i, separating z from the external boundary arc opposite v_i. In particular, **for z lying near v_4, $s_\delta^1(z)$ is almost exactly the crossing probability for the original 4-marked domain.**

The discrete derivatives of the s_δ^i satisfy a $2\pi/3$-rotational version of the Cauchy-Riemann equations. As a result it can be shown that the s_δ^i (extended by interpolation to continuous functions on $\overline{D}$) converge uniformly to a **"harmonic conjugate triple"** (s^1, s^2, s^3) satisfying a mixed Dirichlet problem on D, which has a unique and conformally invariant solution.

An extremely detailed proof of Smirnov's theorem, with full justification for the approximation by discrete domains, is presented in [BR].

8.4 Conclusion

In fact, crossing probabilities are only the simplest possible example of a conformal invariant. Smirnov also proves the much more general result that the **"full percolation configuration"** is conformally invariant in the scaling limit: If we encode a particular configuration by a collection of curves (for example, by the external perimeters of connected components), then the probability laws of these curves are conformally invariant in the scaling limit of the lattice. This shows for instance that the **percolation interface**, depicted in Figures 8.3 and 8.4, follows a conformally invariant probability law in the scaling limit just as Brownian motion does.

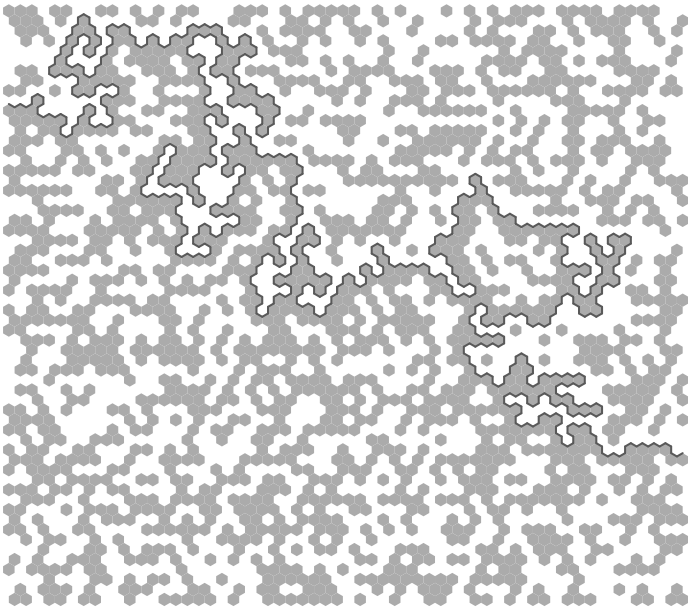


Figure 8.3: A portion of the percolation interface

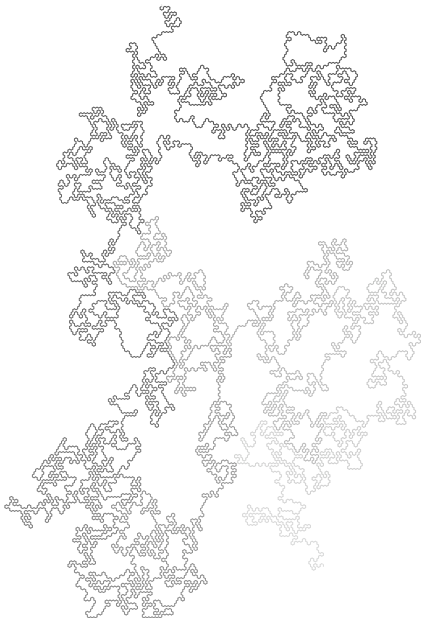


Figure 8.4: At a smaller mesh

Proving this result requires more technical work on the convergence of probability laws for random curves, so we refer the interested reader to the paper by Aizenman and Burchard [AB], and to Smirnov's original paper. We note only that this result differs significantly from Lévy's result for Brownian motion because the curve is **self-avoiding**,[1] and therefore does not have independent increments, so the result has surprisingly strong implications. For example, if γ is a conformally invariant self-avoiding **stochastic process** which starts on the boundary of a simply connected domain D and travels in D, by the Riemann mapping theorem there is a map φ which takes $D \setminus \gamma[0,t]$ conformally onto D, with γ_t mapped by the continuous extension to γ_0. The curve $\varphi(\gamma[t,\infty))$ then follows the *same* law that the original curve γ followed. This observation is a starting point for the theory of **Schramm-Loewner evolutions**; for further reading see [Sc, We].

8.5 Acknowledgements

This topic was first introduced to me by Professor Yum-Tong Siu, and many others have helped me by suggesting useful references and teaching me the relevant background material. I am especially grateful to Professor Wilfried Schmid for answering my questions on complex analysis, for many helpful conversations on this topic, and for his comments on a draft of this article.

References

[Ah] Lars V. Ahlfors: *Complex analysis*, 3rd ed. New York: McGraw-Hill 1979 (International Series in Pure and Applied Mathematics).

[1] More precisely, the curve has no transversal self-intersections. It is possible for it to have double points in the scaling limit.

[AB] Michael Aizenman and Almut Burchard: Hölder regularity and dimension bounds for random curves, *Duke Math. J.* **99** #3 (1999), 419–453.

[BR] Béla Bollobás and Oliver Riordan: *Percolation*. Cambridge: Cambridge Univ. Press 2006.

[Ga] Jean-François Le Gall: *Some properties of planar brownian motion*. Berlin: Springer 1992, (*Lecture Notes in Math.* **1527**), 111–235.

[Gr] Geoffrey Grimmett: *Percolation*, 2nd ed. New York: Springer 1999, (*Grundlehren der Mathematischen Wissenschaften* **321**).

[Ka] Shizuo Kakutani: Two-dimensional brownian motion and harmonic functions, *Proc. Imp. Acad. Tokyo* **20** (1944), 706–714.

[LPSA] Robert Langlands, Philippe Pouliot, and Yvan Saint-Aubin: Conformal invariance in two-dimensional percolation, *Bull. Amer. Math. Soc.* **30** #1 (1944), 1–61.

[La] Gregory F. Lawler: *Conformally Invariant Processes in the Plane*. Providence, RI: American Mathematical Society 2005, (*Math. Surveys and Monographs* **114**).

[Le] Paul Lévy: *Processus stochastiques et mouvement brownien*, 2nd ed. Paris: Gauthier-Villars Paris 1965.

[RW] L. C. G. Rogers and D. Williams: *Diffusions, markov processes and martingales*, 2nd ed., vol. 1. Cambridge: Cambridge Univ. Press 2001.

[Ro] Jeffrey S. Rosenthal: *A first look at rigorous probability theory*, 2nd ed. Hackensack, NJ: World Scientific 2006.

[Sa] Muhammad Sahimi: *Applications of Percolation Theory*. Bristol, PA: Taylor & Francis Inc. 1994.

[Sc] Oded Schramm: Conformally invariant scaling limits, to appear in ICM 2006 Madrid Proceedings (2006).

[Sm] Stanislav Smirnov: Critical percolation in the plane, `http://www.math.kth.se/~stas/papers/`.

[Sm] Stanislav Smirnov: Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits, *C. R. Acad. Sci. Paris Sér. I Math.* **333** (2001), 239–244.

[St] J. Michael Steele: *Stochastic calculus and financial applications*. New York: Springer 2000.

[SS] Elias M. Stein and Rami Shakarchi: *Complex analysis*, Princeton: Princeton Univ. Press 2003.

[Va] S. R. S. Varadhan: *Stochastic processes*. New York: American Mathematical Society 2007 (*Courant Lecture Notes in Mathematics* **16**).

[We] Wendelin Werner: *Random planar curves and Schramm-Loewner evolutions*. New York: Springer 2004 (Lecture Notes in Mathematics, Lectures on Probability Theory and Statistics, *Ecole d'Été de Probabilités de Saint-Flour* XXXII-2002 (Jean Picard, ed.)).

STUDENT ARTICLE

9

APPLIED MATHEMATICS CORNER

DNA Computation and Algorithm Design

Shrenik Shah[†]
Harvard University '09
Cambridge, MA 02138
sshah@fas.harvard.edu

9.1 Introduction

DNA computing was developed by a team run by Leonard Adleman in a 1994 experiment [Ad]. Since then, scientists have produced a number of developments in this area, both theoretical and practical. Adelman's interest in DNA computing arose out of an effort to harness the power of the massive parallelism present in biological systems; theoretical computer science has developed parallel algorithms that efficiently speed up deterministic computation.

The key benefit of DNA computation is **parallelism**—a large number of processes may be run simultaneously. According to [KGY], just a liter of weak DNA solution can hold 10^{19} bits of information, which can encode the states of 10^{18} processors. All of these processors can be acted on simultaneously by the primitive operations of DNA computation. Of course, the structure of this information requires innovative approaches to details that are taken for granted when working with parallel computer grids. When encoding a problem as DNA, it becomes difficult to control the computation and extract the output.

DNA computation has been demonstrated in increasingly more powerful trials. In 2000, Yoshida and Suyama demonstrated a protocol for solving instances of 3-SATISFIABILITY, a very important computational problem [YS]. The instance they solved was rather simple, however—a human could work it out without a computer. A team led by Adleman, who originally developed DNA computing, managed to solve a 20-variable instance of this problem in 2002 [BCJ$^+$]. This difficult instance is beyond human capacity, though this is still just a couple seconds' work for a modern computer.

The status of DNA computation today is still tentative. DNA computation has not yet exceeded the power of modern computers. There are several issues, including the expense of the equipment necessary and the error rate inherent to biological processes, that may prove to cripple the feasability and utility of DNA computation in the long run.

9.2 The DNA Computation Model

To put DNA computation on a concrete framework, the article [Ka] by Lila Kari breaks the process into smaller steps that can be regarded as primitive operations for the DNA computer.

1. **Synthesizing**: This stage involves creating a single DNA strand consisting of polynomially many base pairs.

2. **Mixing**: This step involves taking the contents of two test tubes and mixing them together in a third. This step may seem frivolous, but becomes important in the theoretical framework.

[†]Shrenik Shah, Harvard '09, is a senior mathematics concentrator with a secondary field in English and is also pursuing a concurrent masters' degree in Computer Science. He was a founding member and Articles Editor of The HCMR. His interests lie in algebraic number theory and complexity theory.

3. **Annealing**: This process, also known as **hybridization**, involves lowering the temperature of the solution so that two DNA sequences to attach together in a double helix.

4. **Amplifying**: A reaction known as the Polymerase Chain Reaction (PCR) allows one to produce a copy of a DNA strand. In [Ka], Kari observes that exponentially many strands can be produced by repeatedly performing this operation in parallel.

5. **Separating**: In this step, gel electrophoresis is used to determine the length of a DNA strand.

6. **Extracting**: A step called **affinity purification** allows one to find strands that match a given substring of DNA.

7. **Cutting**: Certain enzymes allow one to cut DNA at sites with particular patterns of bases.

8. **Ligating**: This is the opposite of cutting; certain enzymes provide for the ability to connect DNA strands with certain endings.

9. **Substituting**: This fairly complex operation allows for insertions, deletions, or substitutions of sequences of base pairs in a DNA strand.

10. **Detecting and Reading**: Once a DNA sequence is present in solution, this stage involves determining the sequence of base pairs that compose that strand of DNA in order.

These processes are standard laboratory procedures used by biologists in performing genetic analyses. It is fortunate that these operations are completely parallelizable.

9.2.1 Cutting, Ligating, and Substitution

The processes of **cutting**, **ligating** and **substitution** are frequently used together to patch together DNA sequences. The concatenation of two sequences, an operation that is very frequently used in DNA computation, is actually a sequence of a ligation, a cut, a ligation, a cut, and one simple final step, as described in [KGY]. Both cutting and ligating use the same enzymes that organisms use for the maintenance of their own DNA. For example, cutting uses a restriction endonuclease ([Ka]). The substitution operation is even more complex and requires more steps. From the perspective of algorithm design, one should regard cutting, ligating, and substitution as the main tools for "string manipulation" on DNA.

9.2.2 Amplification and the Polymerase Chain Reaction

Cells in the body need to replicate their DNA on a regular basis, and use the enzyme **DNA polymerase** to do this. The **Polymerase Chain Reaction** (PCR) repeats this process of replication many times, using the new strands created by the replication to produce many new strands in parallel. Through this process, one can acheive exponential growth: one strand becomes two, which becomes four, and so on. This process, called amplification, thus rapidly produces a very large number of copies of the original DNA. It occurs in a solution containing chemicals from which base pairs are constructed, as DNA polymerase cannot create new strands from nothing. At the end of the amplification, then, the DNA must be separated out from this solution. The actual PCR requires a series of heating and cooling cycles, and due to its prevalence in biological research, the lengths and nature of these cycles have been carefully optimized [JK].

9.2.3 Separation and Gel Electrophoresis

Gel electrophoresis is a technique that separates a solution of different DNA strands by length. Detailed accounts of this procedure are found in [LBM$^+$] and [PRS]. A prepared solution of DNA strands having known lengths is used as a kind of "ruler" for comparison with the DNA placed in other wells, and an electric current is passed through a gel in which solutions of DNA strands are placed in wells on one end. The DNA travels slowly through the gel, moved by the electromotive force, with shorter strands traveling more quickly than the longer ones due to a negative charge on the phosphate group of every nucleotide. When the current stops, so does the DNA. In this

way, the DNA is separated by size, which can be measured by comparison to the "ruler." In order to see where the strands ended, one stains the molecules with ethidium bromide, which flouresces under ultraviolet light. Alternatively, one can attach radioactive labels to the DNA and use radiation screening techniques.

9.2.4 Extraction and Affinity Purification

In order to find a known DNA sequence, one can use a complementary strand—called a **probe** in [PRS]—to "fish" for that sequence. A more complicated procedure can even allow one to search for two strands in different solutions that match each other, using a method described in [KGY]. The idea behind more general examples of extraction/affinity purification is to use the natural property that DNA binds to its complement to search for sequences identical to or similar to a known sequence. One can, for example, use this technique progressively, testing larger and larger strands to sequence the DNA, in a manner described in [PRS]. From the perspective of algorithm design, one should use these techniques to test for the existence of a string in a solution.

9.3 DNA Algorithms

In order to illustrate the ways in which the operations discussed in Section 9.2 can be combined to carry out a computation, we present an example, the Bounded Post Correspondence Problem (BPCP). Our focus will be to show two aspects of this process:

- How a problem framed in computational terms can be translated into statements about DNA strands, in a method that lends itself naturally to using those strands for computation. There are many important considerations here, including, for example, the use of complementarity to "search" for a particular sequence in a solution of DNA molecules.

- How the primitive operations fit together and can be used to obtain a much more powerful procedure than originally imaginable.

The algorithm discussed will address a very important computational problem. The Bounded Post Correspondence Problem is classified as $\mathbf{NP}$-complete. There are no known efficient algorithms to solve these problems on a standard computer. The parallelism in DNA computing will become very concrete in this example.

We first define a few terms that will be essential in what follows. We may use a letter such as u to stand for a sequence of base pairs, such as $GCCTA$. Given two sequences, u and v, the **concatenation** $u \cdot v$, usually written simply as uv, is just the concatenation of these two sequences of base pairs. For example, if $u = TA$ and $v = GC$, $uv = TAGC$. Concatenation is made possible using the primitive operation of ligating.

It will be important to encode numbers as well, since these are essential to most computations. For this, we use some kind of precursor sequence followed by the number in base 4, with, say, $A = 0$, $C = 1$, $G = 2$, $T = 3$. It is also possible to write the number in binary using just, say, A and T. The details of this encoding are unimportant in what follows; what is important is that each number has a unique, known encoding.

We next introduce the computational problem, translated into the language of DNA. A detailed account of this algorithm may be found in [KGY].

Problem. *Suppose one is given two collections of DNA strands, $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$, where each u_i and v_i may be an arbitrary sequence of base pairs, as well as an integer $K \leq n$. Do there exist integers $i_1, \ldots, i_\ell$, such that $1 \leq i_j \leq n$ for $j = 1, \ldots, \ell$, and $\ell < K$, such that the concatenation $u_{i_1} \ldots u_{i_\ell}$ is the same as the concatenation $v_{i_1} \ldots v_{i_\ell}$?*

Note that the i_j may include repetitions, and that u_i and v_i are not required to have the same length.

Example 1. Let $u = (AGT, AGA, TAG, GAG)$, $v = (AG, AA, TTAG, AGGG)$, and $K = 4$. Then $u_1 \cdot u_3 = AGT \cdot TAG = AGTTAG$, which is the same as $v_1 \cdot v_3 = AG \cdot TTAG = AGTTAG$. Thus the answer is "yes."

Example 2. Let $u = (AAGTATAG, GATATCC, AGTA, CCAA)$ and $v = (AA, TA, GTA, A)$. Then every single strand of u is longer than every strand of v, so it is impossible that for some choices of i_j, $u_{i_1} \dots u_{i_\ell}$ can match $v_{i_1} \dots v_{i_\ell}$, since the first of these is a longer sequence.

The algorithm to solve this problem is as follows:

1. *Synthesizing Needed Strands*: One needs to produce DNA strands for each u_i and v_i, as well as strands encoding the integers i. For reasons that will become clear momentarily, one should choose an additional sequence we will denote $\#$ that is easily recognizable and not a substring of some concatenation of the u_i and v_i. The sequence $\#$ should also be made into DNA strands. The $\#$ strands will act as markers in strings of base pairs to separate information. Such a strand is called a **bridge** by [KGY]. The $\#$ stands should be placed into two test tubes we denote by U and V. Finally, set a counter k to be 1.

2. *Creating a Solution of Concatenations*: The following routine will be repeated, incrementing k each time:

   We pour the contents of U into n test tubes $U_1, \dots, U_n$, and similarly, V into test tubes $V_1, \dots, V_n$. Recall that this operation is called **mixing**, and is one of the primitive operations in Section 9.2. Then, for each test tube U_i, we prepend u_i to the beginning of every DNA strand inside, and append i to the end of every DNA strand inside. The same is done for each test tube V_i. Finally, we mix the contents of the U_i back into U and the contents of the V_i back into V.

   When $k = 1$, the result is a number of strands of the form $u_{i_1} \# i_1$ in U, where $1 \leq i_1 \leq n$, and similarly strands of the form $v_{i_1} \# i_1$ in V. After repeating this process for k steps, one obtains strands of the form $u_{i_1} \dots u_{i_\ell} \# i_\ell \dots i_1$, where $1 \leq i_j \leq n$ for $j = 1, \dots, \ell$, and similarly for V. Note that we have listed the indices $i_1, \dots, i_\ell$ in increasing order, but u_{i_ℓ} is in fact the first strand prepended to $\#$.

3. *Checking for Matching Strings*: After each run of the routine in Step 2, before repeating the routine, one checks for matching strings in U and V. This can be done by first using affinity purification. If a match is found, the algorithm outputs "yes" and halts. Otherwise, the value of k is incremented. If $k = K$, the algorithm outputs "no" and halts.

9.4 Algorithm Design

The algorithm in Section 9.3 suggests some general principles regarding DNA algorithm design. We discuss the relationship of the BPCP to a class of problems called $\mathbf{NP}$, and then propose a general approach to algorithm design for problems within this class.

9.4.1 Remarks on NP Computation

The class $\mathbf{P}$ contains, roughly, the computational problems a computer can solve when restricted to a polynomial number of time steps, where the polynomial is viewed as a function of the input size. For example, multiplication is such a computation, and the standard multiplication technique constitutes a polynomial-time multiplication algorithm.

The class $\mathbf{NP}$ consists of problems whose answers are easy to check, but for which it may be difficult to come up with an answer. For example, one such problem is that of determining, given a list of linear inequalities in variables x_i, whether there exists a solution in integers to all of these inequalities. It is very easy to check that a proposed solution satisfies the inequalities, but may be very difficult to determine whether one exists. A working solution is called **witness** to the solvability of the problem. Some problems, including the Bounded Post Correspondence Problem, are **complete**, and it is known that if one $\mathbf{NP}$-complete problem can be solved in polynomial time, then *every* problem in $\mathbf{NP}$ also has a polynomial-time algorithm.

The most important characterization of the class $\mathbf{NP}$ for our purposes is as follows: Any problem in $\mathbf{NP}$ may be solved by testing all strings of some bounded length n (a set that is exponential in size) using a single polynomial-time algorithm in hopes of finding a solution. In fact, a general

result from complexity theory shows that this algorithm can be completely parallelized. In the example provided in Section 9.3, the algorithm produced a search space in the first two stages of the algorithm, and tested them in the third.

Another elegant algorithm to solve an **NP**-complete problem, SATISFIABILITY, is presented in [BDLS]. Thus, assuming that the computational problem lies in **NP**, we may divide the computation into the two processes of creating a test space and testing each strand in this space.

9.4.2 Creating a Test Space

In the computation above, all of the new strands at each stage are prepared at the same time, so the growth in the total number of strands is exponential in the number of steps. Given a sufficient quantity of DNA at the outset, one could potentially solve very large instances of the Bounded Post Correspondence Problem.

In general, it is relatively easy to generate all strings of some length by using a concatenation technique repeatedly. However, it can be fruitful to create a limited search space, all of whose members represent "likely" witnesses to the search problem. In the algorithm presented in Section 9.3, this restriction was to search for strings obtained by concatenating members from one of the two collections u and v.

9.4.3 Searching the Test Space

One next needs a technique to determine the result of the computation. Since we now have a space of test strings, we need to check, for each one, whether it is a witness for the problem. If, for example, one simply needed to check for the presence of a particular string, one might use try to match the solution with the complement of a desired strand—this is the extraction technique discussed in Section 9.2.4. One could also use a more complicated method, described in [KGY], in order to check whether there are identical strands in a pair of given solutions.

More generally, one can construct a parallelizable test for whether a witness is correct. The general procedure is somewhat technical, so we merely sketch the idea. If the top node of the circuit is an `OR` gate, then (inductively) one creates a test space S where the strings in S each are complementary strings, forming a "test" for the conditions leading into the `OR` gate. One also creates a second test space T of all strings of some length, and any string that binds to an element of S upon mixing is acceptable. If the node is a `NOT` gate, where one is trying to take the complement of a set S of strings, one can create a test space T of all strings of some length, and any string that fails to bind to an element of S is considered acceptable. In this case, the resulting strings that are left constitute the complement of the desired set, but one can just permute the alphabet to continue.

9.4.4 Procedural Efficiency

The algorithm in Section 9.3 is particularly efficient because it requires few steps. In general situations, one might benefit by recasting a computational problem in a manner amenable to easy searching of the test space, because in practice this process could become excessively complicated.

9.5 Limitations

DNA computation, as the above discussion reveals, has an enormous potential to speed up important computations by a factor of 10,000 or more. However, there are limitations to DNA computation that, while not crippling, suggest that it may not be very useful in practice.

9.5.1 Sequence length

It is very difficult to synthesize sequences much longer than 100–200 base pairs, so either DNA computation must restrict itself to using short strands, or better synthesis technology must be developed. Moreover, longer strands are less easily distinguishable by gel electrophoresis, more prone to error in amplification, and more likely to denature under slight stresses. This is a limitation that has an analogy in memory limitation with computers, but is far more severe. These errors can be crippling to many computing applications, so this could potentially threaten the ability of DNA computation to be useful for certain computational problems even if they are parallelizable. That said, for medium-sized problems, meaning problems that would take a grid of computers on the order of a month to 10 years to solve, DNA computing appears to be a very promising approach.

9.5.2 Error rate

Enzymes inherently have their own error rate, and papers such as [KGY] make special modifications to their algorithms to provide error-correction, usually using additional enzymes that serve this purpose in actual organisms. It may become the case that certain potential applications of DNA computation, which cannot tolerate even a small percentage chance of error, will fail for this reason. On the other hand, the technique of repetition and taking a majority vote of the result can always shrink a constant-sized error rate to become exponentially small, albeit at the cost of additional material use. Since the computations can be done in parallel, the repetition might not involve too substantial an increase in computing time. Researchers have even found certain error-resistant strands that could be used, for example, as the bridge in the algorithm described above. ([Ba] is a patent on some such strands.)

9.5.3 Resources

Working with DNA is very expensive, and certain procedures that seem simple in theory require a great amount of time and care to perform without errors. For a computation with $n = 100$, one would need to purchase at least 300 strands at a cost of $7,500 and wait for a lab to produce the strands. If a problem required running several such algorithms, the costs could very quickly become unmanagable. It may be that only fully automated processes such as those discussed in [RWB$^+$] will be able to cost-effectively carry out practical DNA computing algorithms.

9.6 Conclusions

DNA computation, while still in its infancy, could potentially be a new source of computing power. The largest existing parallel grids contain less than a million computers, while a single liter of DNA solution can hold the states of 10^{18} simple processors. If we could eliminate the major contributors to the cost of DNA contribution, both in time and money, by coming up with easily automated mechanisms for each type of primitive operation, then DNA computation could potentially outrun these grids, particularly on computations involving exponential-sized search spaces.

This automation of the procedures for DNA manipulation necessary for DNA computing could also help biology and chemistry researchers, who often face fairly repetitive work, a large amount of time and energy. If DNA computation ever becomes economically viable, the investments that could pour in from companies who specialize in engineering could make some great developments along the lines of mechanization as well, which could then be used for biological and chemical research.

Neither previous work on algorithms nor on parallel computation harness the full power of DNA computation, since the primitive operations for these models are so different. Algorithm design involves an interesting new set of tradeoffs between space, time, and money. The last of these considerations does not usually enter into standard algorithm design, but is frequently important when working with DNA.

References

[Ad] Leonard M. Adleman: Molecular computation of solutions to combinatorial problems, *Sci.* **266** (1994), 1021–1024.

[Ba] E. B. Baum: DNA sequences useful for computation, US Patent 6,124,444 (2000).

[BCJ$^+$] R. S. Braich, N. Chelyapov, C. Johnson, P. W. K. Rothemund, and L. Adleman: Solution of a 20-Variable 3-SAT Problem on a DNA Computer, *Sci.* **296** #5567 (2002), 499.

[BDLS] Dan Boneh, Christopher Dunworth, Richard J. Lipton, and Jiri Sgall: On the computational power of DNA, *DAMATH: Disc. App. Math. & Comb. Op. Res. & Comp. Sci.* **71** (1996).

[JK] N. Jonoska and S. A. Karl: Ligation experiments in computing with DNA, *IEEE International Conference on Evol. Comp.* (1997), 261–266.

[Ka] Lila Kari: DNA computing: arrival of biological mathematics, *The Math. Intelligencer* **19** #2 (1997), 9–22.

[KGY] Lila Kari, Greg Gloor, and Sheng Yu: Using DNA to solve the Bounded Post Correspondence Problem, *Theor. Comp. Sci.* **231** #2 (2000), 193–203.

[LBM$^+$] H. Lodish, A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, M. P. Scott, S. L. Zipursky, and J. Darnell: *Molecular Cell Biology*, 5th ed. New York: W. H. Freeman and Co. 2003.

[PRS] G. Paun, G. Rozenberg, and A. Salomaa: *DNA Computing: New Computing Paradigms*. Berlin: Springer 1998.

[RWB$^+$] S. T. Roweis, E. Winfree, R. Burgoyne, N. V. Chelyapov, M. F. Goodman, P. W. K. Rothemund, and L. M. Adleman: A Sticker-Based Model for DNA Computation, *J. of Comp. Bio.* **5** #4 (1998), 615–630.

[YS] H. Yoshida and A. Suyamaf: Solution to 3-SAT by Breadth First Search, *DIMACS Workshop DNA Based Computers* #5 Massachusetts Institute of Technology (2000).

FEATURE

10

MY FAVORITE PROBLEM

An Unconventional Inequality

Ameya Velingker[†]
Harvard University '10
Cambridge, MA 02138
avelingk@fas.harvard.edu

10.1 Introduction

Problem solving has been an important aspect of mathematics in my life. It is the challenge of tackling a math problem and experiencing the moments of both insight and perplexity associated with the problem solving process that have drawn me to the field of mathematics.

When asked to write about my "favorite" problem, I found it rather difficult to single out a particular problem from my repository of interesting mathematics problems. Though I do not have a single favorite problem, I tried to choose a problem which meets several criteria that make a mathematics problem interesting. First, the problem should have a simple/elegant solution; at the same time, the solution should have some key step which is clever and difficult. The ideal problem is the one that looks intractable upon first sight but, after one has read the solution, should evoke the response, "Ah, that was simple! Why didn't I think of that?" Moreover, the problem should have interesting generalizations or connections to other problems or ideas in mathematics.

Keeping in mind these characteristics of a good problem, I have selected one which I believe meets the criteria. But first, I will present some background on the problem. The problem was created by Reid Barton and submitted as a problem for the 2003 International Mathematical Olympiad (IMO). Though it was not selected as a question on the IMO, it was included on the IMO Shortlist, an annual list of twenty to thirty problems which are in contention for a place on the IMO exam. The problem was first presented to me during the 2004 USA Mathematical Olympiad Summer Program (MOSP), where it was given as a problem on a practice test. As I recall, no one was able to solve it, and I was quite fascinated after learning its clever solution.

10.2 Problem

The problem is stated below:

Problem. *Let n be a positive integer and let $(x_1, \ldots, x_n)$, $(y_1, \ldots, y_n)$ be two sequences of positive real numbers. Suppose $(z_2, \ldots, z_{2n})$ is a sequence of positive real numbers such that*

$$z_{i+j}^2 \geq x_i y_j \text{ for all } 1 \leq i, j \leq n. \tag{10.1}$$

Let $M = \max\{z_2, \ldots, z_{2n}\}$. Prove that

$$\left(\frac{M + z_2 + \cdots + z_{2n}}{2n}\right)^2 \geq \left(\frac{x_1 + \cdots + x_n}{n}\right)\left(\frac{y_1 + \cdots + y_n}{n}\right). \tag{10.2}$$

The first thing that strikes the reader about this problem is how unconventional it is. The condition given in (10.1) is certainly bizarre, as is the appearance of M in the inequality. Proving

[†]Ameya Velingker, Harvard '10, is a mathematics and physics concentrator living in Currier House. His academic interests span a wide range of subjects, including number theory, analysis, mathematical physics, and algorithms. Outside of academics, he enjoys tennis, pool, and Indian classical music.

inequalities is a common type of question in the olympiad exams, and any experienced olympiad problem solver has in his arsenal a number of tools to attack such questions, such as AM-GM inequality, Cauchy-Schwarz inequality, and Muirhead's inequality (see [HLP]), to name a few. The trouble with this inequality is that none of the standard tricks seem to work, as we will highlight.

The first questions that emerge regarding the inequality are how strict it is and what equality cases, if any, there are. Upon quick inspection, one notices that choosing $x_1 = x_2 = \cdots = x_n = y_1 = y_2 = \cdots = y_n = z_2 = z_3 = \cdots = z_{2n}$ satisfies (10.1) and yields equality in our inequality. This equality condition, combined with the nature of the left-hand side of (10.2), is reminiscent of the Inequality of Arithmetic and Geometric Means (or AM-GM, for short):

Theorem 1 (AM-GM)**.** *Given a list of k nonnegative real numbers $a_1, a_2, \ldots, a_k$, the following inequality holds:*

$$\frac{a_1 + a_2 + \cdots + a_k}{k} \geq \sqrt[k]{a_1 a_2 \ldots a_k},$$

with equality if and only if $a_1 = a_2 = \cdots = a_k$.

A simple proof of this inequality can be found in ([HLP]).

Since the equality case of (10.2) appears to require $M = z_2 = z_3 = \cdots = z_n$, we are tempted to try applying the AM-GM inequality on the left-hand side of (10.2):

$$\frac{M + z_2 + z_3 + \cdots + z_{2n}}{2n} \geq \sqrt[2n]{M z_2 z_3 \ldots z_{2n}}.$$

Thus, proving (10.2) reduces to showing that

$$\sqrt[n]{M z_2 z_3 \ldots z_{2n}} \geq \left(\frac{x_1 + \cdots + x_n}{n}\right)\left(\frac{y_1 + \cdots + y_n}{n}\right).$$

However, the problem is that the above inequality is actually false, in general. One can come up with numerous counterexamples; for instance, take $n = 3$ with $x_1 = y_1 = 1$, $x_2 = y_2 = 2$, $x_3 = y_3 = 3$, $z_2 = 1$, $z_3 = \sqrt{2}$, $z_4 = 2$, $z_5 = \sqrt{6}$, and $z_6 = 3$, and the above inequality is not satisfied. Thus, this approach does not work, as AM-GM is too weak for the left-hand side.

Another approach is to use the Cauchy-Schwarz inequality, which in sequence form, states the following:

Theorem 2 (Cauchy-Schwarz Inequality)**.** *Given $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n \in \mathbb{R}$, we have*

$$\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right) \geq \left(\sum_{i=1}^{n} x_i y_i\right)^2.$$

The trouble is that the only apparent candidate for application of Cauchy-Schwarz would be the right side of (10.2), in the form of

$$\left(\frac{x_1 + \cdots + x_n}{n}\right)\left(\frac{y_1 + \cdots + y_n}{n}\right) \geq \left(\frac{\sqrt{x_1 y_1}}{n} + \cdots + \frac{\sqrt{x_n y_n}}{n}\right)^2.$$

However, this inequality goes in the wrong direction, so we are forced to abandon this idea.

Another idea uses a different application of AM-GM, this time on the right side of (10.2). The right side is a product of two quantities and lends itself to AM-GM:

$$\begin{aligned}\left(\frac{x_1 + \cdots + x_n}{n}\right)\left(\frac{y_1 + \cdots + y_n}{n}\right) &\leq \left(\frac{\frac{x_1+\cdots+x_n}{n} + \frac{y_1+\cdots+y_n}{n}}{2}\right)^2 \\ &= \left(\frac{x_1 + \cdots + x_n + y_1 + \cdots + y_n}{2n}\right)^2.\end{aligned}$$

In light of the above inequality, it suffices to establish

$$M + z_2 + z_3 + \cdots + z_{2n} \geq x_1 + \cdots + x_n + y_1 + \cdots + y_n,$$

which seems to be a simpler inequality. The reader can try some examples and convince himself that the above inequality appears to be true. Thus, this line of attack seems promising.

However, in attempting to prove (10.2), none of the standard tricks appear to work. The tricky part lies in using the condition (10.1) effectively. One could reasonably expect to use $z_k \geq \sqrt{x_i y_j}$ for some $i + j$ and prove some inequality of the form

$$M + \sum \sqrt{x_i y_j} \geq x_1 + \cdots + x_n + y_1 + \cdots + y_n,$$

where the sum ranges over certain pairs (i, j). However, the direction of the inequality makes it nearly intractable to tackle via inequalities such as AM-GM or Muirhead.

Justifiably so, it is at this juncture that a brilliant maneuver is required. First, we take advantage of the homogeneity of (10.1) and (10.2). Let $x = \max\{x_1, \ldots, x_n\}$ and $y = \max\{y_1, \ldots, y_n\}$. Then without loss of generality, one may replace each x_i with $x_i' = x_i/x$, each y_i with $y_i' = y_i/y$, and each z_i with $z_i' = z_i/\sqrt{xy}$ without affecting the statement of the problem. Thus, it suffices to prove (10.2) under the added assumption that $\max\{x_1, x_2, \ldots, x_n\} = \max\{y_1, y_2, \ldots, y_n\} = 1$.

Now, the critical ingredient in the proof is the following lemma:

Lemma 3. *Let $a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_k$ be positive reals. Suppose that, for any $r > 0$, the following property is satisfied:*

(i). The number of i for which $a_i > r$ is at least the number of i for which $b_i > r$.

Then, $a_1 + \cdots + a_n \geq b_1 + \cdots + b_n$.

Proof. Without loss of generality, one can assume that $a_1 \leq a_2 \leq \cdots \leq a_k$ and $b_1 \leq b_2 \leq \cdots \leq b_k$. Note that if there exists k for which $b_k > a_k$, then, the number of i for which $b_i > (a_k + b_k)/2$ is at least $n - k + 1$ (since $i = k + 1, \ldots, n$ satisfy the relation), while the number of i for which $a_i > (a_k + b_k)/2$ is at most $n - k$, contradicting the initial assumption. Hence, we must have $a_i \geq b_i$ for all i, and so, $a_1 + \cdots + a_n \geq b_1 + \cdots + b_n$. □

The lemma seems rather obvious, but it is powerful enough to establish (10.2) and thus provide a complete solution.

10.3 Solution

We now present a complete solution (provided by [DJMP]) to the problem along these lines:

Proof. Let $x = \max\{x_1, \ldots, x_n\}$ and $y = \max\{y_1, \ldots, y_n\}$. Then, without loss of generality, we may assume that $x = y = 1$, for we can always replace x_i by x_i/x, y_i by y_i/y, and z_i by $z_i/\sqrt{xy}$ without affecting the statement of the problem. It suffices to show that

$$M + z_2 + \cdots + z_{2n} \geq (x_1 + \cdots + x_n) + (y_1 + \cdots + y_n)$$

because applying the AM-GM inequality to the sum of two terms on the right side of the above inequality would give us the desired result.

Now, by the previous lemma, we need only show that, for any $r > 0$, the number of terms on the left side of (10.2) that are greater than r is at least the number of those on the right side. Observe that if $r \geq 1$, this property clearly holds, as no terms on the right side are greater than r.

Next, suppose $r < 1$. Let $X = \{i : x_i > r\}$, $Y = \{i : y_i > r\}$, and $Z = \{i : z_i > r\}$. Note that if $x_i, y_j > r$, then $z_{i+j} \geq \sqrt{x_i y_j} > r$. Thus,

$$\{i + j : i \in X, j \in Y\} \subseteq Z.$$

However, note that X and Y are nonempty (since $r < 1$ and $x = y = 1$). Thus, if $X = \{a_1, a_2, \ldots, a_l\}$ and $Y = \{b_1, b_2, \ldots, b_m\}$ with $a_1 < a_2 < \cdots < a_l$ and $b_1 < b_2 < \cdots < b_m$,

then $\{a_1 + b_1, a_1 + b_2, \ldots, a_1 + b_m, a_2 + b_m, a_3 + b_m, \ldots, a_m + b_m\} \subseteq Z$, which shows that $|Z| \geq |X| + |Y| - 1$. But then, we also have that $M > r$. Hence, there are at least $|X| + |Y|$ elements on the left side of (10.2) that are greater than r.

This concludes the proof. □

The proof is everything we want it to be: short, elegant, and clever.

10.4 Further Connections

What I find remarkable about the problem, apart from the simple (albeit difficult) nature of its solution, is the multitude of connections it has with the field of discrete geometry. As it turns out, the inequality we have discussed is related to an important inequality known as the Prékopa-Leindler Inequality:

Theorem 4 (Prékopa-Leindler Inequality)**.** *Let $0 < \lambda < 1$, and let $f, g, h : \mathbb{R}^n \to [0, \infty)$ be measurable functions such that*

$$h(\lambda x + (1 - \lambda)y) \geq f(x)^{\lambda} g(y)^{1-\lambda}$$

for all $x, y \in \mathbb{R}^n$. Then,

$$\int_{\mathbb{R}^n} h(x)\, dx \geq \left(\int_{\mathbb{R}^n} f(x)\, dx \right)^{\lambda} \left(\int_{\mathbb{R}^n} g(x)\, dx \right)^{1-\lambda}.$$

Of course, the most striking difference between the given problem and the statement of the Prékopa-Leindler Inequality is the fact that the former deals with sequences while the latter deals with functions. In fact, our inequality can be viewed as a discrete specialization of the Prékopa-Leindler inequality for $n = 1$ and $\lambda = 1/2$.

The Prékopa-Leindler inequality has many important applications, such as probability theory, optimal mass transportation [Vi], and the theory of diffusion [BL]. Perhaps its most important consequence is the Brunn-Minkowski inequality (see [Ba]), which can be stated as follows (note that there exist alternative formulations of the inequality):

Theorem 5. *If A and B are compact subsets of $\mathbb{R}^n$, then*

$$|\lambda A + (1 - \lambda)B|^{1/n} \geq \lambda |A|^{1/n} + (1 - \lambda)|B|^{1/n},$$

where $|X|$ denotes the Lebesgue measure of X, and $\lambda A + (1 - \lambda)B$ denotes the Minkowski sum $\{\lambda a + (1 - \lambda)b : a \in A, b \in B\}$.

The Brunn-Minkowski inequality can be used to provide a simple proof (from [Ba]) of the famous isoperimetric inequality:

Theorem 6 (Isoperimetric Inequality)**.** *Among simple closed bodies of a given volume in $\mathbb{R}^n$, Euclidean balls have the least surface area.*

Proof. Let $C \in \mathbb{R}^n$ be a compact set with volume equal to that of B_n, the Euclidean ball of radius 1. Then, the surface area of C is given by

$$|\partial C| = \lim_{\epsilon \to 0} \frac{|C + \epsilon B_n| - |C|}{\epsilon}.$$

By the Brunn-Minkowski inequality, we have

$$|C + \epsilon B_n| \geq \left(|C|^{1/n} + \epsilon |B_n|^{1/n} \right)^n \geq |C| + n\epsilon |C|^{(n-1)/n} |B_n|^{1/n}.$$

It then follows that

$$\begin{aligned} |\partial C| &\geq n|C|^{(n-1)/n} |B_n|^{1/n} \\ &= n|B_n|. \end{aligned}$$

Using the well-known fact $n|B_n| = |\partial B_n|$ (see page 4 of [Ba]), we obtain $|\partial C| \geq |\partial B_n|$, as desired. □

References

[Ba] K. M. Ball: An elementary introduction to modern convex geometry (1997).

[BL] H. J. Brascamp and E. H. Lieb: On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation, *J. Functional Analysis* **22** (1976).

[DJMP] D. Djukić, V. Janković, I. Matić, and N. Petrović: *The IMO Compendium*. New York: Springer Science+Business Media, Inc. (2006).

[HLP] G. H. Hardy, J. E. Littlewood, and G. Polya: *Inequalities*, Cambridge: Cambridge Univ. Press (1952).

[Vi] C. Villani: *Topics in Optimal Transportation*, American Mathematical Society (2003) (*Grad. Stud. in Math.* **58**).

FEATURE

11 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. 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, as well as 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 *April 13, 2009*; any problems not solved in this admittedly short window will be reopened in Vol. 3, No. 1 with a solution submission deadline of September 21, 2009. An asterisk beside a problem or part of a problem indicates that no solution is currently available.

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).

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

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

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).

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).

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).

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 $\angle BID = \angle CIE = \angle AIF = 90^\circ$, 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).

The following two problems from the Spring 2008 issue are being released for one more issue. The first required correction and clarification, and we are grateful to Daniel Kane, G2 for bringing these issues to our attention. The second problem below 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, Paul Kominers (MIT '12), and Justin Chen (Caltech '09).

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).

FEATURE

12
Solutions

Seeing Stars

F07 – 5. For $i = 1, \ldots, n$, let $f_i : (\mathbb{Z}/m\mathbb{Z} \cup \{\star\})^n \to (\mathbb{Z}/m\mathbb{Z} \cup \{\star\})^n$ be given by

$$f_i\left((x_1, \ldots, x_n)\right) = \begin{cases} (\star, x_2 + 1, x_3, \ldots, x_n) & i = 1 \text{ and } x_1 = 1, \\ (x_1, \ldots, x_{i-1} + 1, \star, x_{i+1} + 1, \ldots, x_n) & 1 < i < n \text{ and } x_i = 1, \\ (x_1, \ldots, x_{n-2}, x_{n-1} + 1, \star) & i = n \text{ and } x_n = 1, \\ (x_1, \ldots, x_n) & \text{otherwise,} \end{cases}$$

where $\star + 1 = \star$. Find necessary and sufficient conditions on $(x_1, \ldots, x_n) \in (\mathbb{Z}/m\mathbb{Z})^n$ such that there exists a sequence $\{i_k\}_{k=1}^n$ for which

$$f_{i_n}(\cdots(f_{i_1}((x_1, \ldots, x_n)))) = (\star, \ldots, \star).$$

Proposed by Paul Kominers (Walt Whitman HS '08), Scott D. Kominers '09, and Zachary Abel '10.

Solution by Benjamin Dozier '12. Call an n-tuple $(x_1, \ldots, x_n) \in (\mathbb{Z}/m\mathbb{Z})^n$ **starrable** if and only if there exists a sequence $\{i_k\}_{k=1}^n$ with

$$(f_{i_n} \circ \cdots \circ f_{i_1})(x_1, \ldots, x_n) = (\star, \ldots, \star). \tag{12.1}$$

We call any sequence $\{i_k\}_{k=1}^n$ for which (12.1) holds a **starring sequence for $(x_1, \ldots, x_n)$**.

Call an n-tuple $(x_1, \ldots, x_n)$ **uninull** if and only if $\sum_{j=1}^k x_j = 0$ or 1 for $1 \le k \le n-1$ and $\sum_{j=1}^n x_j = 1$. We claim that the starrable n-tuples are precisely those that are uninull.

If $m = 1$ then the only n-tuple is $(0, \ldots, 0)$, which is both starrable and uninull. For the rest of the proof, we assume $m > 1$.

We prove the claim by induction on n. For $n = 1$, the only 1-tuple that is starrable is (1), which is also the only 1-tuple that is uninull. Assume the claim for n. Let $A = (x_1, \ldots, x_{n+1})$ be a starrable $(n+1)$-tuple with starring sequence $\{i_k\}_{k=1}^{n+1}$. First note that

(i) initially none of the elements of A are stars,

(ii) they all become stars after we have applied a starring sequence of functions, and

(iii) if f_j changes the ith element of an n-tuple from a non-star to a star then $i = j$.

We conclude that the set of all elements of the starring sequence equals the set $\{1, 2, \ldots, n\}$. In particular, all of the elements of the starring sequence are distinct.

Now if x_1 is not equal to 0 or 1, then $(f_{i_n} \circ \cdots \circ f_{i_1})(A)$ will have first element not equal to $\star$, since the only functions that can change the first element are f_1 and f_2, but f_2 either increments the first element by 1 or does not affect the first element and f_1 changes the first element from 1 to $\star$. This is a contradiction since A is starrable and $\{i_k\}_{k=1}^{n+1}$ is a starring sequence for A; thus x_1 is equal to either 1 or 0.

If $x_1 = 0$ it is easy to see that 2 must precede 1 in the sequence $\{i_k\}$ since f_1 only outputs a sequence that has a star as the first element if the first element of the input is 1 or $\star$ and f_2 must have already been applied for this to be the case. Since f_1 affects at most two elements of its input $(n+1)$-tuple, the first and the second, and since f_2 is applied before f_1, changing the second

element to a star which will remain a star even after f_1 is applied, we conclude that f_1 only affects the first element of the $(n+1)$-tuple. Thus $(x_2, \ldots, x_{n+1})$ must be a starrable n-tuple, and, by the inductive hypothesis, also uninull. But then $(0, x_2, \ldots, x_{n+1})$ is uninull.

Now we consider the case when $x_1 = 1$. It is easy to see that 2 must come after 1 in the sequence $\{i_k\}$ because f_1 and f_2 are the only functions that can affect the first element, but if f_2 is applied before f_1 then the first element becomes $2 \neq 1$, and thus f_1 will have no effect. Thus the effect of f_1 will be to change the first element to a star and increase the second element by 1. It then follows that $(x_2 - 1, x_3, \ldots, x_{n+1})$ must be a starrable n-tuple, and, by the inductive hypothesis, also uninull. But then $(1, x_2, x_3, \ldots, x_{n+1})$ is uninull.

Combining the $x_1 = 0$ and $x_1 = 1$ cases we see that any starrable $n+1$ tuple must also be uninull. Now we prove the converse.

Let $(a_1, \ldots, a_n)$ be a starrable n-tuple in $(\mathbb{Z}/m\mathbb{Z})^n$ with starring sequence $i'_1, \ldots, i'_n$. Then $(0, a_1, \ldots, a_n)$ is a starrable $(n+1)$-tuple with starring sequence $i'_1, \ldots, i'_n, 1$. Also, $(1, a_1 - 1, \ldots, a_n)$ is a starrable $(n+1)$-tuple with starring sequence $1, i'_1, \ldots, i'_n$.

Now note that if an n-tuple $(x_1, \ldots, x_n)$ is uninull then x_1 is either 0 or 1. If the former holds, then $(x_2, \ldots, x_n)$ is also uninull and thus starrable by the inductive hypothesis. But then, as discussed above, $(x_1, \ldots, x_n)$ is starrable. Alternatively, if $x_1 = 1$ then $(x_2 + 1, x_3, \ldots, x_n)$ is uninull—thus starrable by the inductive hypothesis—and $(x_1, \ldots, x_n)$ is again starrable by the previous paragraph. □

Also solved by Kenfin Tomioka (University of Tokyo, Japan) and the proposers.

Symmetrized Sudoku Kernels

S08 – 1. It is known that there are 6670903752021072936960 square matrices M of order 9 with entries in $\{1, \ldots, 9\}$ that show valid sudoku grids.[1] How many of them have the property that the symmetric matrix $M + M^t$ is positive definite?

Proposed by Noam D. Elkies (Harvard University).

Solution by the proposer. We show that there are 0 such matrices. We prove this by showing that every such matrix satisfies $w(M + M^t)w^t = 0$ where w is the nonzero vector

$$w = (1, 1, 1, -1, -1, -1, 0, 0, 0).$$

Let $e_1, \ldots, e_9$ be the standard unit vectors in $\mathbb{R}^9$, and for each of $j = 1, 2, 3$ let $v_j = e_{3j-2} + e_{3j-1} + e_{3j}$, so

$$\begin{aligned} v_1 &= (1, 1, 1, 0, 0, 0, 0, 0, 0), \\ v_2 &= (0, 0, 0, 1, 1, 1, 0, 0, 0), \\ v_3 &= (0, 0, 0, 0, 0, 0, 1, 1, 1). \end{aligned}$$

Then for $j, k \in \{1, 2, 3\}$ we have

$$v_j M v_k^t = v_j M^t v_k^t = \sum_{i=1}^{9} i = 45,$$

because $v_j M v_k^t$ and $v_j M^t v_k^t$ are the sum of the entries in the (j, k)-th and (k, j)-th 3×3 block of the Sudoku array M. It follows that the vector $w = v_1 - v_2$ satisfies $wMw^t = wM^tw^t = 0$, whence $w(M + M^t)w^t = 0$ as claimed.

[1] The proposer points out that this calculation is detailed in Bertram Felgenhauer and Frazer Jarvis: Enumerating possible Sudoku grids (2005), `http://www.afjarvis.staff.shef.ac.uk/sudoku/sudoku.pdf`, athough it was independently computed by user "QSCGZ" on the rec.puzzle Google group, thread "combinatorial question on 9x9," 21 Sep. 2003.

Remark. We could have used for w any nonzero vector in the 2-dimensional space

$$V = \{a_1v_1 + a_2v_2 + a_3v_3 \mid a_1 + a_2 + a_3 = 0\}.$$

It follows that if $M + M^t$ is positive *semi*definite then its kernel contains V. We have not found any such M, but neither can we prove that none exists. □

Also solved by Daniel Kane, G2.

Diophantine Squeeze

S08 – 3. Let $k \geq 1$ be a natural number. Find all integer solutions to the diophantine equation

$$x^{2k+1} + x^{2k} + \cdots + x^2 + x + 1 = y^{2k+1}.$$

Proposed by Ovidiu Furdui (University of Toledo).

Solution by the Missouri State University Problem Solving Group. Clearly, $(x, y) = (0, 1)$ or $(-1, 0)$ are solutions for all k. We claim that there are no other solutions. Note that

$$\sum_{i=0}^{2k} x^i = \begin{cases} (x^{2k+1} - 1)/(x - 1) & \text{if } x \neq 1 \\ 2k + 1 & \text{if } x = 1 \end{cases}$$

which is positive for all x. Therefore

$$x^{2k+1} < \sum_{i=0}^{2k+1} x^i \text{ for all } x.$$

If $x > 0$, then

$$x^{2k+1} < \sum_{i=0}^{2k+1} x^i = y^{2k+1} < \sum_{i=0}^{2k+1} \binom{2k+1}{i} x^i = (x+1)^{2k+1}$$

which is clearly impossible for integers x and y.

If $x < -1$, we will prove that $(-1)^n \sum_{i=0}^{n} x^i > (-1)^n (x+1)^n$ for $n \geq 2$ by induction on n. Since $x < 0$, this is clearly true when $n = 2$. Assume that the result holds when $n = m$, *i.e.*

$$(-1)^m \frac{x^{m+1} - 1}{x - 1} = (-1)^m \sum_{i=0}^{m} x^i > (-1)^m (x+1)^m.$$

Multiplying both sides by $-(x + 1)$ (which is positive), we obtain

$$(-1)^{m+1} \frac{x^{m+2} + x^{m+1} - x - 1}{x - 1} > (-1)^{m+1} (x+1)^{m+1}.$$

Now since $x < -1$, $(-1)^{m+1}(x^{m+1} - x)/(x - 1)$ is negative regardless of whether m is even or odd, so

$$\begin{aligned} (-1)^{m+1} \sum_{i=0}^{m+1} x^i &= (-1)^{m+1} \frac{x^{m+2} - 1}{x - 1} \\ &> (-1)^{m+1} \frac{x^{m+2} + x^{m+1} - x - 1}{x - 1} \\ &> (-1)^{m+1} (x+1)^{m+1} \end{aligned}$$

which is what we needed to show.

We are interested in the case $n = 2k + 1$ where the inequality we just proved becomes

$$\sum_{i=0}^{2k+1} x^i < (x+1)^{2k+1}.$$

As in the case when $x > 0$, we have

$$x^{2k+1} < \sum_{i=0}^{2k+1} x^i = y^{2k+1} < (x+1)^{2k+1}$$

which is again impossible. □

Also solved by the Northwestern University Problem Solving Group, Koichiro Nomura (University of Tokyo, Japan) and the proposer.

E. Equilateralibus Isosceles

S08 – 5. Let ABC be a non-isosceles triangle with $\angle A = 60^\circ$. Let H be its orthocenter and I its incenter. Let B_i and C_i the points such that the equilateral triangles ABC_i and AB_iC intersect the interior of ABC. Define B_e and C_e similarly, so that ABC_e and AB_eC are equilateral and disjoint from the interior of ABC.

Show that the lines through HI, B_iC_i and B_eC_e do not concur, and that the triangle they form is isosceles.

Proposed by Daniel Campos Salas (Costa Rica).

Solution by Yasuhide Minoda (Tetsuryokukai Institute, Japan). Without loss of generality, we may assume $AB < AC$. Let $X = B_iC_i \cap AB_e$. Since $\angle CBC_i = \angle BC_iA - \angle BCC_i = \frac{\pi}{3} - \angle C$ and

$$\begin{aligned} \angle CBC_i &= \angle CB_iC_i \quad \text{(because } B_i, B, C_i, C \text{ are concyclic)} \\ &= \angle B_iXA \quad \text{(because } B_iC \text{ and } AX \text{ are parallel),} \end{aligned}$$

we have $\angle B_iXA = \frac{\pi}{3} - \angle C$.

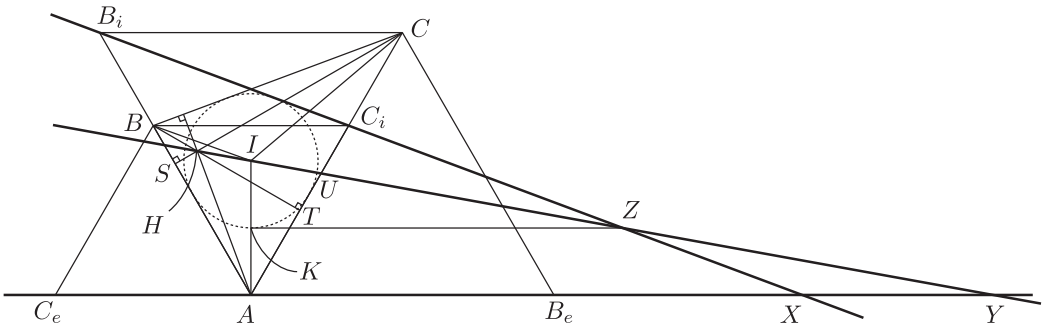


Figure 12.1: Diagram for Problem S08–5.

Next, let $Y = HI \cap AB_e$; we now show $\angle IYA = \frac{1}{2}\angle B_iXA$. Let $S = CH \cap AB$ and $T = BH \cap AC$. We have $\angle BHC = \angle SHT = 2\pi - \frac{\pi}{2} - \frac{\pi}{2} - \angle A = \frac{2\pi}{3}$ by considering quadrilateral $HSAT$, and from triangle IBC it follows that

$$\angle BIC = \pi - (\angle IBC + \angle ICB) = \pi - \frac{1}{2}(\angle B + \angle C) = \pi - \frac{1}{2}(\pi - \angle A) = \frac{2\pi}{3}.$$

It follows that $\angle BHC = \angle BIC$, from which we conclude (letting $U = HI \cap AC$):

$$\begin{aligned}
&B, H, I, C \text{ are concyclic} \\
&\implies \angle IHT = \angle ICB = \frac{1}{2}\angle C \\
&\implies \angle CUY = \angle HUT = \frac{\pi}{2} - \angle IHT = \frac{\pi}{2} - \frac{1}{2}\angle C \\
&\implies \angle IYA = \angle CUY - \angle UAY = \frac{\pi}{2} - \frac{1}{2}\angle C - \frac{\pi}{6} = \frac{\pi}{6} - \frac{1}{2}\angle C \\
&\implies IYA = \frac{1}{2}\angle B_iXA.
\end{aligned}$$

Therefore, if HI, B_iC_i, and B_eC_e do not concur, the triangle they form is isosceles.

Now we have to show that HI, B_iC_i and B_eC_e do not concur. Draw a tangent line to the incircle of triangle ABC from $Z = HI \cap B_iC_i$, other than ZB_i, and let K be the tangent point. Clearly $\angle B_iZI = \angle IZK$. Since $\angle B_iZI = \angle IYA$, we have $\angle IZK = \angle IYA$. Thus ZK and AY are parallel, so K is the intersection of the incircle of triangle ABC and the segment IA. It is clear that $K \neq A$, so we have $Z \neq X$. This means that HI, B_iC_i, and B_eC_e do not concur. □

Also solved by Koichiro Nomura (University of Tokyo, Japan) and the proposer.

FEATURE

13

ENDPAPER

Hunting for Perfect Euler Bricks

Oliver Knill[†]
Harvard University
Cambridge, MA 02138
knill@math.harvard.edu

An **Euler brick** is a cuboid with integer side dimensions a, b, c such that the face diagonals are integers. The cuboid with dimensions $(a, b, c) = (44, 117, 240)$, for example, is an Euler brick. It is the smallest Euler brick. If (a, b, c) is an Euler brick, then (ka, kb, kc) is an Euler brick too for positive integers k. If also the space diagonal is an integer, an Euler brick is called a **perfect Euler brick**. In other words, a perfect Euler brick has the properties that all vertex coordinates and vertex distances are integers.

Whether a perfect Euler brick exists is an open mathematical problem. One would have to find integer vectors (a, b, c) such that

$$\sqrt{a^2 + b^2}, \sqrt{a^2 + c^2}, \sqrt{b^2 + c^2}, \sqrt{a^2 + b^2 + c^2}$$

are integers. Nobody has found a solution to this system of Diophantine equations nor shown that solutions do not exist. A infinite subclass of Euler bricks can be parametrized: if u, v, w is a Pythagorean triple $u^2 + v^2 = w^2$, then

$$(a, b, c) = (|u(4v^2 - w^2)|, |v(4u^2 - w^2)|, |4uvw|)$$

is an Euler brick.

Because $a^2 + b^2 + c^2 = f(t, s)(s^2 + t^2)^2$ if $u = 2st; v = s^2 - t^2; w = s^2 + t^2$ and $f(t, s) = s^8 + 68s^6t^2 - 122s^4t^4 + 68s^2t^6 + t^8$, it would suffice to find s, t for which $f(t, s)$ is a square in order to find a perfect Euler brick. There are many Euler bricks which do not fall into the above Saunderson parametrization known since 1740. A brute force search $1 \leq a \leq b \leq c \leq 8000$ leads to 120 Euler bricks. Only 16 of the 120 Euler bricks in $1 \leq a \leq b \leq c \leq 8000$ are prime bricks, triples (a, b, c) which are not a multiple of a smaller brick. Some of them, like $(85, 132, 720)$ are not of the above parametrization. To look for perfect Euler bricks of the parametrized type we can search for integers $\sqrt{f(t, s)}$ with the help of a computer. Since perfect Euler bricks might not exist, one can try to find Euler bricks (a, b, c) for which $\sqrt{a^2 + b^2 + c^2}$ is as close to an integer as possible. One approach is to linearize the map $T : \sqrt{f(t, s)} \to \sqrt{f(t + u, s + v)} \mod 1$ for suitable (u, v) and use a continued fraction expansion of the irrational rotation $dT(x) = x + \alpha \mod 1$ on $[0, 1)$ to find n for which $dT^n(x)$ is close to 0. There exists for example a number a with 68162 digits, a number b with 56802 digits and a number c with 56803 digits so that the diagonal length $\sqrt{a^2 + b^2 + c^2}$ is 10^{-60589} close to an integer. Computations with such large numbers push the boundaries of computer algebra systems. One has to take square roots of integers with hundreds of thousands of digits. It turns out that in such ranges, some computer algebra systems have limitations when projecting algebraic numbers to real valued numbers. While the quest for Euler bricks might appear an entertainment without applications, the treasure hunt can at least help to explore the boundaries and limitations of computer algebra systems.

[†]Oliver Knill got his mathematics degrees at ETH Zuerich in Switzerland and has been in California, Arizona, and Texas before coming to Harvard in 2000. He especially likes to think about mathematical topics which are accessible to undergraduates and which also allow exploration, visualization, and experimentation with the help of computers.