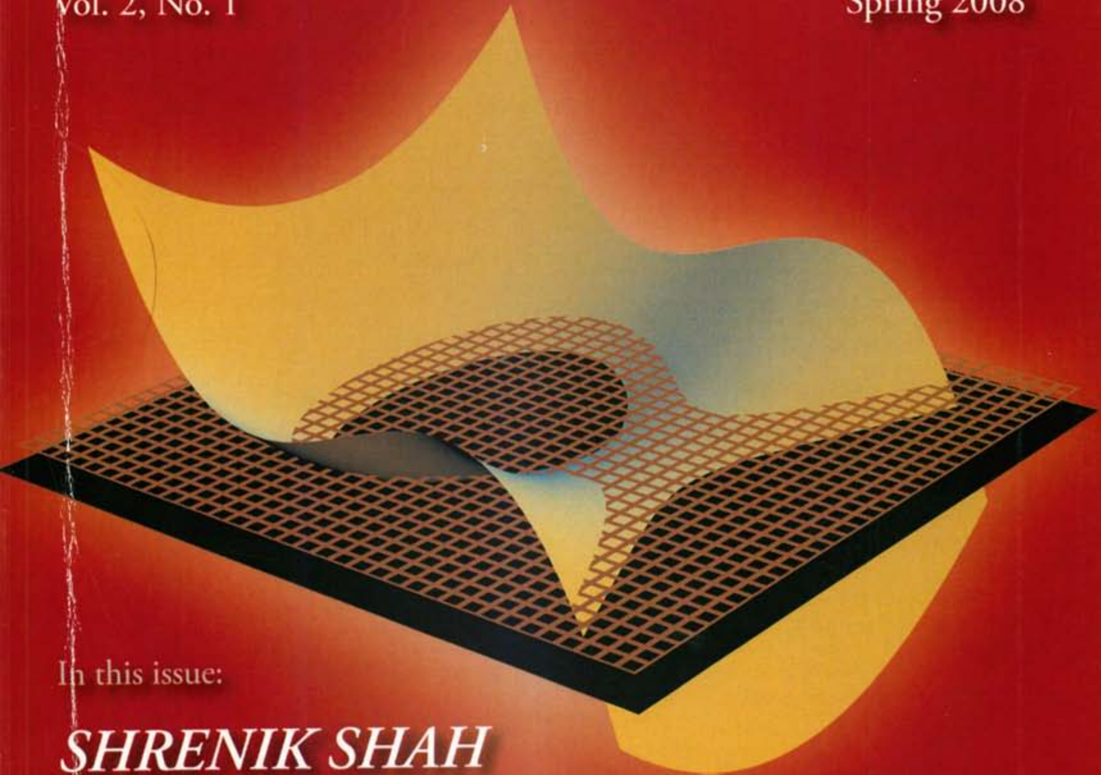
The Harvard College Mathematics Review

Vol. 2, No. 1

Spring 2008



In this issue:

SHRENIK SHAH

A Taste of Elliptic Curve Cryptography

THOMAS LAM

Tiling with Commutative Rings

DANA ROWLAND

Twisting with Fibonacci

HC MR

A Student Publication of Harvard College

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

$$\texttt{http://www.thehcmr.org/} \qquad (1)$$

Instructions for Authors. All submissions should include the name(s) of the author(s), institutional affiliations (if any), and both postal and e-mail addresses at which the corresponding author may be reached. General questions should be addressed to Editor-In-Chief Scott D. Kominers 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 Charles Nathanson.

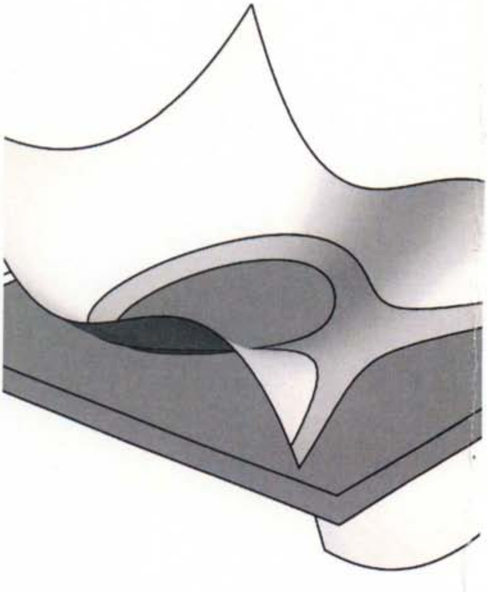
Subscriptions. One-year (two issue) subscriptions are available, at rates of $10.00 for students, $15.00 for other individuals, and $30.00 for institutions. Subscribers should mail checks for the appropriate amount to The HCMR's postal address; confirmation e-mails should be directed to Distribution Manager Nike Sun, at 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 +

Fellows · Dr. Barbara Currier · Ellen and William Kominers · Teach for America · **Contributors** · The Harvard Undergraduate Council · QVT Financial LP · **Patrons** · The Harvard University Mathematics Department

Cover Image. The image on the cover shows a surface whose level sets are real elliptic curves. An application of elliptic curves is the subject of this issue's "A Taste of Elliptic Curve Cryptography," by Shrenik Shah (p. 13). The image was created in Mathematica$^{\text{TM}}$ by Graphic Artist Zachary Abel.



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

-2
Contents

Student Articles

Faculty Feature Article

Features

-1
Staff

0
From the Editor

Scott D. Kominers
Harvard University '09
Cambridge, MA 02138
kominers@fas.harvard.edu

This spring, Professor John Duncan, the instructor teaching my algebraic geometry class, assigned supplementary readings from *The Harvard College Mathematics Review* (HCMR):

"Alexander Ellis's article gives an excellent taste of the Poincaré-Hopf Index Theorem,"

Professor Duncan explained, strongly recommending that all the students read Ellis's article (Dunking donuts: Culinary calculations of the Euler characteristic, *HCMR* **1** #1 (2007), 3–14) to complement the day's lecture.

In a single year, The HCMR has already reached the status of a teaching tool. Indeed, articles from The HCMR have appeared not only in Harvard classrooms but also in classes at other schools (*e.g.* Professor Ivars Peterson's "Communicating Mathematics" course) and in summer programs (*e.g.* Professor Keith Conrad's 2007 PROMYS lectures).

This new step in The HCMR's maturation is truly impressive. The HCMR could never have reached this level without its scholarly **authors**, nor could have it even existed without its dedicated **editors**, tireless **business staff**, and sleepless **production directors**.

Furthermore, the success of The HCMR has been facilitated by the journal's **sponsors** and **advisers**, both at Harvard and elsewhere. In particular, we at The HCMR are especially indebted to **Professor Benedict H. Gross '71**, **Professor Peter Kronheimer**, and **Dr. Alon Amit**, who have advised the journal and to **Professor Clifford H. Taubes** for his continued support and encouragement. We also appreciate the administrative assistance provided by **Dean Paul J. McLoughlin II** and **Mr. David R. Friedrich** as well as the continued, generous support of **The Harvard Mathematics Department**.

This issue is my last as Editor-In-Chief. I am glad to see that The HCMR's future looks so bright at this juncture.

The incoming Co-Editors-In-Chief, **Zachary Abel '10** and **Ernest E. Fontes '10**, are both dedicated members of The HCMR's staff. Both are talented mathematicians and energetic leaders, and both have been deeply involved in the production of this and past issues. I look forward to seeing The HCMR continue to expand under their joint administration.

On a personal note, I would like to thank those teachers most responsible for my involvement in the founding of The HCMR, {**Mrs. Susan Schwartz Wildstrom**, **Professor Noam D. Elkies**, **Professor Andreea C. Nicoara**}, and to wish good luck to the new Co-Editors-In-Chief. $\mathbb{QED}$.

Scott D. Kominers '09
Editor-In-Chief, The HCMR

STUDENT ARTICLE

1 Latin Squares and their Partial Transversals

Nikolaos Rapanos†
Kastritsi High School '08
Patras, Greece
nrapanos@gmail.com

Abstract

We introduce the theory of Latin squares and their partial transversals. Furthermore, we present the history and some applications of Latin squares. We also prove using elementary methods that a 6×6 Latin square has a partial transversal of length 5. After developing some of the theory of so-called partial Latin squares we provide a simple proof of a result originally due to Woolbright, which gives a lower bound for the length of the longest partial transversal in an $n \times n$ Latin square. Followingly we improve slightly the best known lower bound for the length of the longest partial transversal in an $n \times n$ Latin square.

1.1 Introduction

A **Latin square** of order n is an $n \times n$ square matrix whose cells consist of n distinct symbols such that each symbol appears exactly once in each row and each column. The theory of these objects has a long history. The earliest written reference of Latin squares is the solution to the card problem published in 1723 (see [DK] for further information). Euler (1779) initiated the systematic development of the theory, when he posed "The Problem of the 36 Officers" in his paper "Recherches sur une nouvelle espece de quarre magique".[1] Later, Cayley (1877-1890) showed that the multiplication table of a group is a certain type of a Latin square. Latin squares played an important role in the foundation of finite geometries, a subject which was also in development at the turn of the nineteenth century. In the 1930s, a major application of Latin squares was found by Fisher who used them and other combinatorial structures in the design of statistical experiments. Latin squares also find applications in computer science and in the construction of error-correcting telegraph codes.

More on the history and applications of Latin squares can be found in Section 1.2 below. In Section 1.3 we define the concept of a partial transversal of length j of an $n \times n$ Latin square. We introduce an operation # which will enable us to prove some existence results for partial transversals. We then survey some lower bounds on the length of the longest partial transversal in an $n \times n$ Latin square. In Section 1.4 we define partial Latin squares and use some of their properties to prove one of the aforementioned lower bounds. We conclude by sketching how to improve on the proven bound.

1.2 History and Applications of Latin Squares

Latin squares have a long history, stretching back at least as far as medieval Islam, when they were used on amulets. Abu l'Abbas al Buni wrote about them and constructed, for example, 4×4 Latin

†Nikolaos Rapanos has a high school senior at Kastritsi High School in Patras, Greece. He is member of his National Math team during the last 6 years and he has been distinguished in various math competitions, such as the IMO-BMO. His mathematical interests include Analysis, Geometry, and Inequalities. Beyond math, his interests include physics, music, tennis and waterskiing. He plans to study Engineering, although a double major in mathematics is increasingly likely.

[1] English: Investigations on a new species of magic square.

squares using letters from a name of God. In his famous etching Melancholia, the fifteenth century artist Albrecht Dürer portrays a 4 × 4 magic square, a relative of Latin squares, in the background. Other early references to them concern the problem of placing the 16 face cards of an ordinary playing deck in the form of a square so that no row, column, or diagonal should contain more than one card of each suit and each rank. This is known as "The Card Problem."

The systematic treatment of Latin squares started by Leonhard Euler in 1779, when he posed "The Problem of the 36 Officers." This problem was not solved until the beginning of the twentieth century. The problem was to arrange 36 officers, each having one of six different ranks and belonging to one of six different regiments, in a square formation 6 × 6, so that each row and each file contained just one officer of each rank and just one from each regiment. Working on this problem, Euler defined Graeco-Latin—or **orthogonal**—Latin squares as a pair of $n \times n$ Latin squares so that when one is superimposed on the other, each of the n^2 combinations of the symbols (taking the order of the superimposition into account) occurs exactly once in the n^2 cells of the array. "The Problem of the 36 Officers" can be solved by finding a pair of orthogonal Latin squares of order 6. Euler knew (c. 1780) that there was not a Graeco-Latin square of order 2 and knew constructions of Graeco-Latin squares for n odd and for n divisible by 4, but he was unable either to find a pair of orthogonal Latin squares of order 6 or prove that they did not exist. Based on much experimentation, he conjectured that Graeco-Latin squares did not exist for $n = 2 \pmod 4$ (oddly even integers according to his notation).

In 1901, Gaston Tarry proved (by exhaustive enumeration of the possible cases) that there exists no Graeco-Latin square of order 6, adding evidence to Euler's conjecture. However, in 1959, Parker, Bose and Shrikhande were able to construct an order 10 Graeco-Latin square, and provided a construction for the remaining even values of n that are not divisible by 4 (of course, excepting $n = 2$ and $n = 6$).

Many more applications of Latin squares were developed in the 20th century. Arthur Cayley continued work on Latin squares and in the 1930s the concept arose again in the guise of multiplication tables when the theory of quasi-groups and loops began to be developed as a generalization of the group concept.

Latin squares played an important role in the foundation of finite geometries, a subject which was also in development at this time. A large application area for Latin squares was opened by Fisher who used them and other combinatorial structures in the design of statistical experiments. Sets of Latin squares that are orthogonal to each other have found an application as error-correcting codes in situations where communication is disturbed by noise besides simple white noise, such as when attempting to transmit broadband internet over powerlines. Latin squares are also useful for scheduling round-robin tournaments. As a matching procedure, Latin squares relate to problems in **graph theory, job assignment** (known as the Marriage Problem), and processor scheduling for massively parallel computer systems. More about applications of Latin squares can be found in [DK].

An interesting open problem seeks a formula $L(n)$ for the number of $n \times n$ Latin squares. The number of $n \times n$ Latin squares is known to increase very fast—in fact $L(n) \geq \prod_{j=1}^{n} j!$—but the exact formula is not yet known.

1.3 Partial Transversals

1.3.1 Definitions

Define a **partial transversal** of an $n \times n$ Latin square to be a set of n cells, one from each row and column. A partial transversal has length j if it contains j distinct symbols. For $j = n$ the partial transversal is called a **transversal**.

We continue with the proof of an elementary fact as an example.

Proposition 1. *Every* 6 × 6 *Latin square has a partial transversal of length at least* 4.

Proof. Assume that there exists a 6 × 6 Latin square whose maximum partial transversal has length less than 4. Then one can find at least 2 same symbols at every partial transversal of the Latin square. Thus, we may assume that our Latin square looks like the one in Figure 1.3.1.

a					
b		x		y	
c					
d	▨	c	▨		▨
e				c	
f					

Figure 1.1: Every 6 × 6 Latin square has a partial transversal of length at least 4.

It is clear that symbols x and y are both different than c and at least one of them is different than a. Without loss of generality, we may examine only the case in which x is different than a, then one of the gray-colored cells, say the (i_0, j_0), must contain a symbol that is neither a nor x. Note that all the gray-colored cells contain symbols that are different than c. Therefore it is clear that the partial transversal consisted of the cells $\{(1,1),(2,3),(5,5),(i_0,j_0)\}$ and two more (acceptable) cells and contains at least 4 distinct symbols—or in other words, its length is at least 4. We work similarly if y is different than a and finish the proof as above. □

1.3.2 The operation

Given an $n \times n$ Latin square and a partial transversal T of length $n - k$ with $k \geq 2$, one can find another partial transversal of equal or greater length in the following manner:

1. Choose two cells (i_1, j_1) and (i_2, j_2) of T such that $T - \{(i_1, j_1), (i_2, j_2)\}$ contains $n - k$ distinct symbols. Since $k \geq 2$, there exist two such cells. These two cells can either contain two distinct duplicated symbols, or two occurrences of the same symbol, if this symbol appears in the transversal at least three times.

2. Replace these two cells with the cells (i_1, j_2) and (i_2, j_1).

Since we chose cells containing duplicated symbols, the new partial transversal has length at least $n - k$, as each of the symbols in the original transversal is represented in one of the unchanged cells. An example of this operation is given in Figure 1.2. We call this operation #, a notation chosen for its shape.

Proposition 2. *Every* 6×6 *Latin square has a partial transversal of length at least 5.*

Proof. This proposition follows directly from Drake's result in [Dr], but we also give another proof as a motivating example for the use of operation #. Let us assume that there is no partial transversal of length 5. There exists a partial transversal of length 4 according to Proposition 1, so there will be a partial transversal containing a multiset[2] of symbols either of the form $\{a, a, b, b, c, d\}$ or the form $\{a, a, a, b, c, d\}$.

Case 3. There is a partial transversal containing symbols of the form $\{a, a, b, b, c, d\}$. An illustration for this case is presented in Figure 1.2.

We may assume without loss of generality that this partial transversal is on the diagonal, and call it T_0. We can apply # to the cells $(1,1)$ and $(3,3)$ in T_0 to get a new partial transversal T_1.

[2] A set whose elements may be repeated is called a **multiset**.

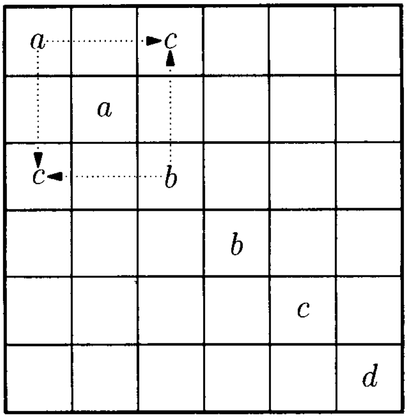


Figure 1.2: An example of the operation #. In this case, we replace the cells $(1, 1)$ and $(3, 3)$ in the partial transversal on the diagonal with the cells $(1, 3)$ and $(3, 1)$ to obtain another partial transversal, also of length 4.

By our hypotheses, the new cells $(1, 3)$ and $(3, 1)$ in T_1 must contain a symbol chosen from the set c, d. Because of symmetry, we only need to analyze two cases: either both symbols are the same or there is one c and one d.

Subcase 3.1. Both symbols obtained from the above application of # are c's.

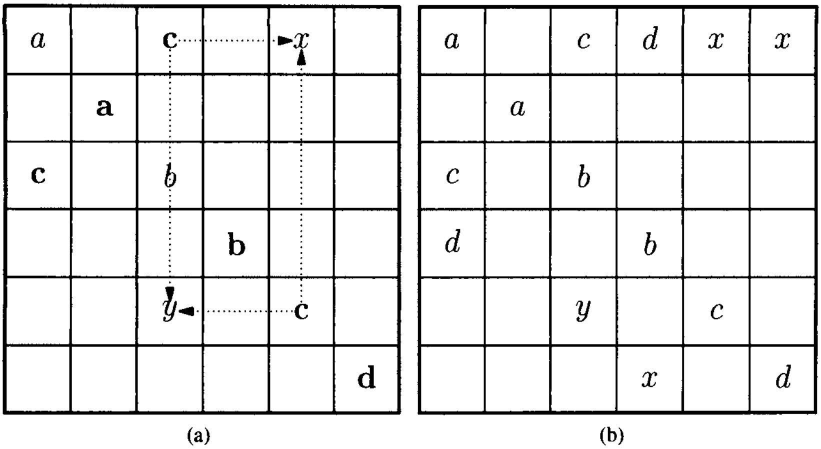
Since both are c's, we can apply # to the cells $(1, 3)$ and $(5, 5)$ in T_1 to obtain a new partial transversal T_2 (as shown in Figure 1.3(a)), and we discover that the symbols in $(1, 5)$ and $(5, 3)$ must be chosen from the set $\{a, b, d\}$. On the other hand, applying # to the cells $(1, 1)$ and $(4, 4)$ in T_0 gives us a partial transversal T_3, and we discover that the symbols in $(1, 4)$ and $(4, 1)$ must both contain d (see Figure 1.3(b)). We can now apply # to the cells $(1, 4)$ and $(6, 6)$ in T_3 to obtain T_4, and discover that the symbols in $(1, 6)$ and $(6, 4)$ must be chosen from the set $\{a, b, c\}$. We know that our Latin square looks like Figure 1.3(b), where the x' s are symbols from the set $\{a, b, c, d\}$. The first row contains five distinct symbols from the set $\{a, b, c, d\}$, a contradiction by the pigeonhole principle.

Subcase 3.2. The symbols of the cells (1,3) and (3,1) are different and chosen from the set $\{c, d\}$.

Without loss of generality, we can assume that the cell $(3, 1)$ contains a c and the cell $(1, 3)$ contains a d. The partial transversal $T_1{}'$ is the one obtained by # to the cells $(1, 1)$ and (3,3) in T_0 . Applying # to the cells (1,1) and (4,4) in T_0, we obtain a new partial transversal T_5. Clearly, the cell (4,1) must be filled with a d and the cell (1,4) with a c. Note, that cells obtained after an application of # cannot contain another symbol, say e, because then we would derive directly our proposition. This is the reason, for which we assume that each cell obtained by # must contain a symbol from the set $\{a, b, c, d\}$. The partial transversal T_6 contains all three c' s of our Latin square and the cells (2,2), (4,3) and (6,6). Applying # to the cells (1,3) and (5,5) in T_6, we obtain a new partial transversal T_7, and we discover that the cell (1,5) must contain a b. Applying # to the cells (6,6) and (1,4), we obtain a new partial transversal T_8 containing the cell (6,1). As we noted above, this cell obtained by # must contain a symbol from the set $\{a, b, c, d\}$. This yields a contradiction, since the first row already contains these symbols.

Case 4. We are left with the case, in which the partial transversal (we may assume it is the diagonal) contains symbols of the form $\{a, a, a, b, c, d\}$.

We may assume, without loss of generality, that the diagonal contains the symbols $\{a, a, a, b, c, d\}$ in this order. Call this partial transversal T_0. Applying # to the cells (1,1) and (2,2) of T_0, we discover that the cells $x_1 = (1, 2)$ and $x_2 = (2, 1)$ must contain symbols of the set $\{b, c, d\}$. Applying # to the cells (1,1) and (3,3) of T_0, we discover that the cells $x_3 = (1, 3)$ and $x_4 = (3, 1)$ must also contain symbols from the set $\{b, c, d\}$. Since $x_1 \neq x_3$, we understand that one of these



a		c		x	
	a				
c		b			
			b		
		y		c	
					d

(a)

a		c	d	x	x
	a				
c		b			
d			b		
		y		c	
			x		d

(b)

Figure 1.3: (a) Another example of the operation #. Here, we replace the cells $(1, 3)$ and $(5, 5)$ in the partial transversal indicated in bold with the cells containing symbols x and y. If this Latin square has no partial transversal of length greater than 4, we must have $x \in \{b, d\}$ and $y \in \{a, d\}$. (b) After two more applications of #, we know that if the Latin square had no partial transversal of length greater than 4, then it would appear as pictured, where the symbols x are chosen from the set $\{a, b, c, d\}$.

two cells contains a different symbol of that in cell x_4. Without loss of generality, we assume that $x_3 \neq x_4$. If we consider the partial transversal containing $\{x_3, x_4, (2, 2), (4, 4), (5, 5), (6, 6)\}$, then we have reduced this case to case 3. □

1.3.3 Open Problems on Latin Squares

Koksma [Ko] showed that an $n \times n$ Latin square has a partial transversal of length at least $\frac{2n+1}{3}$. This was improved by Drake [Dr] to $\frac{3n}{4}$ and then simultaneously by Brouwer et al. [BdVW] and Woolbright [Wo] to $n - \sqrt{n}$. This was in turn improved by Hatami and Shor [HS] to $n - 11.0525(\log n)^2$. We obtain a different lower bound which, while still of order $n - O(\log n)^2$, slightly improves upon Hatami and Shor's constant found in [HS]. Both results fall short of two conjectures, Ryser's conjecture of n for odd n, and Brualdi's Conjecture of $n - 1$. The difficulty of these problems is based on the fact that for large Latin squares, brute force is intractable. In particular, the two conjectures are stated as follows:

Conjecture 5 (Ryser). *Any Latin square of odd order has a transversal.*

Conjecture 6 (Brualdi). *Any Latin square of order n has a partial transversal of length at least $n - 1$.*

1.4 Partial Latin Squares and a Proof of Woolbright's Theorem

We define a **partial Latin square** as an $n \times n$ square matrix of cells, some of which contain a symbol such that no symbol appears twice in any row or column. A **partial transversal** of an $n \times n$ partial Latin square is a set of n filled cells, one from each row and column. We say that a partial transversal of a partial Latin square is of **length** j if it contains j distinct symbols. An $m \times m$ **subsquare** S' of an $n \times n$ partial Latin square S is the set of m^2 cells in some subset of m rows and some subset of m columns of the partial Latin square, where some filled cells of S may possibly be replaced by empty cells in S'.

Consider a Latin square with a partial transversal of maximum length $n - k$, where $k \geq 2$. By applying # to this partial transversal, we obtain other partial transversals, whose lengths must also be $n - k$ and whose set of symbols is the same as the first. Applying # repeatedly to these partial transversals, we eventually obtain a maximal set of partial transversals closed under #. All of these partial transversals contain the same set of $n - k$ distinct symbols, so by ignoring all cells except those in this set of partial transversals, we obtain a partial Latin square S containing $n - k$ distinct symbols and a set $\mathbb{T}$ of partial transversals of S closed under #. We will call this pair $(S, \mathbb{T})$ a partial Latin square satisfying A_k. More formally we have the following definition.

Definition 7. An $n \times n$ **partial Latin square** satisfying A_k is an $n \times n$ array of cells, some of which contain a symbol, together with a nonempty set $\mathbb{T}$ of partial transversals of length $n - k$, that satisfies the following properties:

1. Each filled cell must appear in at least one of the partial transversals in $\mathbb{T}$.

2. The set $\mathbb{T}$ of partial transversals must form a connected graph under #.[3]

3. The set $\mathbb{T}$ of partial transversals must be closed under the operation #.

We consider a partial Latin subsquare $(S', \mathbb{T}')$ satisfying $A_{k'}$ of a partial Latin square $(S, \mathbb{T})$ satisfying A_k. In this case, we require that $\mathbb{T}'$ be a subset of $\mathbb{T}$ restricted to S', i.e. that

$$\mathbb{T}' \subseteq \{T \cap S' : T \in \mathbb{T}\}.$$

Note that the case analysis in the proof of Proposition 2 on Latin squares of size 6 sketched earlier shows that any partial Latin square satisfying A_2 must have size at least 7. Namely, we have already proved that every 6×6 Latin square has a partial transversal of length at least 5.

Lemma 8. *If $(S, \mathbb{T})$ is a partial Latin square satisfying A_k such that no subsquare satisfies A_k, then no cell is contained in all partial transversals in $\mathbb{T}$. In other words, there is no fixed cell. That is, given a filled cell (i, j) and a partial transversal containing (i, j), by a sequence of operations #, one can obtain a partial transversal in $\mathbb{T}$ not containing (i, j).*

Proof. Suppose there is a cell (i, j) contained in all partial transversals. We will call this a fixed cell. Let α be the symbol in this cell. If α appears anywhere else in the partial transversal, then there is a partial transversal containing both α's (the second α appears in some partial transversal since every filled cell does, and this partial transversal must contain the first α since all partial transversals do). We can then apply # to this partial transversal to obtain a partial transversal without the fixed cell, which is a contradiction. We are left with the case in which α does not appear anywhere else in the partial Latin square. By deleting the row and column containing the α, one finds a subsquare satisfying A_k, a contradiction of the hypothesis. □

Proposition 9. *Given a minimal partial Latin square $(S, \mathbb{T})$ satisfying A_k and any filled cell, there exists a partial transversal in $\mathbb{T}$ containing both that cell and another one with the same symbol.*

Proof. We may assume that there is only one cell containing the symbol α. We already know (from Definition 7) that there must be a partial transversal in $\mathbb{T}$ passing through α. Choose a cell on such partial transversal and call it T_1. By Lemma 8 no cell is fixed, so there is a partial transversal T_1 in $\mathbb{T}$ that does not pass through α. On the other hand, the set $\mathbb{T}$ of partial transversals must form a connected graph under the operation #, which is a contradiction. □

Corollary 10. *Given a minimal partial Latin square $(S, \mathbb{T})$ satisfying A_k, there exists a subsquare of S satisfying A_{k-1}.*

Proof. We can choose any filled cell, say (1,1) containing α, and a partial transversal T_0 passing through it that duplicates α. Now consider the set of partial transversals containing at least two α's, including the one in cell (1,1), which are generated by a sequence of operations # starting with T_0. This is exactly the set of partial transversals generated by # starting from the partial transversal $T_0' = T_0 - (1, 1)$ in the subsquare formed by deleting the first row and column. Taking this set of

[3] When we say that the set $\mathbb{T}$ of partial transversals **must form a connected graph under #**, we mean that any transversal in $\mathbb{T}$ can be obtained from any other transversal in $\mathbb{T}$ by repeated applications of #.

partial transversals gives an $(n-1) \times (n-1)$ partial Latin square satisfying A_{k-1}. Note that this subsquare may have some empty cells which were filled in the original $n \times n$ square. □

We now state a lemma, theorem, and corollary proven by Hatami and Shor [HS]:

Lemma 11. *In an $(n-1) \times (n-1)$ partial Latin square satisfying A_{k-1} induced as described above from an $n \times n$ partial Latin square satisfying A_k, the partial transversals generated by # must have a fixed cell,that is some cell that appears in all of these partial transversals.*

Theorem 12. *In a minimal partial Latin square S satisfying A_k, there are at least $n_{k-1} + k$ filled cells in each row and column, where n_{k-1} is the size of the smallest subsquare of S satisfying A_{k-1}.*

Corollary 13. *If we let $n_k = n$ be the size of the original partial Latin square satisfying A_k, then*

$$n_k \geq n_{k-1} + 2k. \tag{1.1}$$

Proof. The larger square has $n_k - k$ distinct symbols, of which at least $n_{k-1} + k$ appear in each row and column. □

1.4.1 An elegant proof of Woolbright's Theorem

Based on our previous analysis, we can give a very short and elegant proof for Woolbright's result [Wo]. Note that the original proof in [Wo] is very complicated.

Let S_k be a partial Latin square satisfying A_k such that no subsquare satisfies A_k. We call such a partial Latin square **minimal**. It was shown in Section 1.4 that there must be a subsquare of S_k satisfying A_{k-1}. Choose S_{k-1} to be the smallest subsquare of S_k satisfying A_{k-1} and, recursively, S_m to be the smallest subsquare of S_{m+1} satisfying A_m, until the sequence ends at S_1. Denote by n_j the size of S_j. In this terminology, Proposition 1 can be expressed as

$$n_2 \geq 7. \tag{1.2}$$

Theorem 14 (Woolbright, 1978). *Every $n \times n$ Latin square has a partial transversal of length at least $n - \sqrt{n}$.*

Proof. Suppose that we are given a $n \times n$ Latin square which does not have a partial transversal of length more than $n - k$. It follows from the definition of n_k, that $n - k \geq n_k - k$. We have already shown that $n_k - n_{k-1} \geq 2k$ in Corollary 13. Adding up, we get:

$$\left.\begin{array}{rcl} n_k - n_{k-1} & \geq & 2 \cdot k \\ n_{k-1} - n_{k-2} & \geq & 2 \cdot (k-1) \\ & \vdots & \\ n_3 - n_2 & \geq & 2 \cdot 3 \\ n_2 - n_1 & \geq & 2 \cdot 2 \end{array}\right\} n_k = 2(1 + 2 + \cdots + k) - 2 = k^2 + k - 2.$$

Thus $n_k \geq k^2$, or in other words $k \leq \sqrt{n_k}$. Consequently, there must be a partial transversal of length at least $n - \sqrt{n}$. □

The reader may enjoy trying to derive the following inequality for the sequence $\{n_k\}$:

Theorem 15. *In S_k as defined above, for all $j < k$,*

$$(n_k - n_j)(n_{k-1} + n_j - n_k + k) \leq n_j(n_j - n_{j-1} - 2j) + (n_k - n_j)(n_k - k - n_j + j). \tag{1.3}$$

Note that inequality (1.3) simplifies to

$$(n_k - n_j)(2n_j + n_{k-1} - 2n_k + 2k - j) \leq n_j(n_j - n_{j-1} - 2j). \tag{1.4}$$

Suppose we have a Latin square, which has no partial transversal of length more than $n - l$. By the previous sections, there is a sequence $n_2 < n_3 < \cdots < n_l$ satisfying the inequalities (1.2) and (1.3) defined earlier. These inequalities hold for $1 \leq j < k \leq l$. We want to place a lower bound on the expression $n - l$. Since $n - l \geq n_k - k$, it is sufficient to place a lower bound on the expression $n_k - k$.

Using the above inequalities we can prove that in an $n \times n$ Latin square, there is a partial transversal of length at least $n - 11.0368(\log n)^2$. Although for large n we have

$$n - 11.0368(\log n)^2 \geq n - \sqrt{n},$$

for small values of n the inequality flips. Finally we can state that every $n \times n$ Latin square has a partial transversal of length at least

$$\max\left\{n - \sqrt{n}, n - 11.0368(\log n)^2\right\}.$$

However we omit the proof of this research since it lies outside of the scope of this article.

Now we give an example of a sequence satisfying inequalities (1.2) and (1.3).

Proposition 16. *The sequence $u_2 = 7$ and $u_k = a^{3\lfloor\sqrt{k}\rfloor}\left(a - b^{k - \lfloor\sqrt{k}\rfloor^2}\right)$ for any $k > 2$, with $a \geq 2^{\frac{1}{3}}$ and $b \leq \frac{1}{4}$, satisfies the inequalities (1.2) and (1.3).*

This proposition shows that inequality (1.3) cannot imply anything better than

$$n - \left(\frac{1}{\log 2}\right)^2 (\log n)^2$$

since the sequence $\{u_k\}$ satisfies all the conditions.

In our research, we only needed to examine the case in which $\{u_k\}$ is an upper bound for the sequence $\{n_k\}$. This is because if there exists an N such that $n_j > u_j$ for all $j \geq N$, then $n_j - j \geq u_j - j$ which gives a much better lower bound for the length of a partial transversal than in the previous case. Furthermore, the slowest growing sequence, that satisfies the inequalities would be a lower bound for the sequence $\{n_k\}$. We were not able to find the slowest growing sequence, but we can show that this sequence is bounded below by $2^{\sqrt{k}}$ and above by

$$\exp\left(\sqrt{k \log \frac{1}{c} \log \frac{4c-2}{c}}\right)$$

for some constant c.

1.5 Acknowledgments

I would like to thank my mentor Dr. Timothy Nguyen for his guidance and his support during the mentorship, Professor David Jerison for overseeing my project, and my tutor Ms. Allison Gilmore for all her assistance. I would also like to thank the Center for Excellence in Education, the Research Science Institute and the Massachusetts Institute of Technology for providing me with the support and opportunity to perform this research. I am thankful to the staff of The HCMR for their assistance and creative cooperation.

References

[BdVW] Andries E. Brouwer, A. J. de Vries, and R. M. A. Wieringa: A lower bound for the length of partial transversals in a Latin square, *Nieuw Arch. Wisk.* **26** #3 (1978), 330–332.

[Br] Victor Bryant: *Aspects of Combinatorics*. Cambridge: Cambridge Univ. Press 1993.

[Ca] Peter J. Cameron: *Combinatorics*. Cambridge: Cambridge Univ. Press 1994.

[DK] Jozsef Dénes and A. Donald Keedwell: *Latin squares and their applications.* New York: Academic Press, 1974.

[Dr] David A. Drake: Maximal sets of Latin squares and partial transversals. *J. Stat. Pl. I.* **1** (1977), 143–149.

[HS] Pooya Hatami and Peter W. Shor: A Lower Bound for the length of a Partial Transversal in a Latin Square, to appear in *J. Combinatorial Theory Ser. A* (2008).

[Ko] K. K. Koksma: A lower bound for the order of a partial transversal in a Latin Square, *J. Combinatorial Theory Ser. A* **7** (1969), 94–95.

[Sh] Peter W. Shor: A lower bound on the length of a partial transversal in a Latin square, *J. Combinatorial Theory Ser. A* **33** #1 (1982), 1–8.

[Wo] David E. Woolbright: An $n \times n$ Latin square has a transversal with at least $n - \sqrt{n}$ distinct elements, *J. Combinatorial Theory Ser. A*, **24** #2 (1978), 235–237.

STUDENT ARTICLE

2
A Taste of Elliptic Curve Cryptography

Shrenik Shah[†]
Harvard University '09
Cambridge, MA 02138
sshah@fas.harvard.edu

Abstract

This paper develops several classical algorithms and cryptosystems in cryptography, and develops the theory of elliptic curves to reveal the improvements provided by elliptic curve cryptography. The prerequisites to this paper are an understanding of groups, fields, and some elementary number theory.

2.1 Introduction

Elliptic curve cryptography is not only a surprising application of a deep and powerful area of number theory to computer science, but as we shall see, is also a very practical technique that is used today around the world. In this article we present a small taste of cryptography, presenting some classical cryptographic protocols and factorization methods. After this we describe some of the mathematical theory of elliptic curves. These are combined in a section that shows how to use elliptic curves in the earlier protocols, and explains why these represent an improvement over traditional methods, while citing some limitations to this viewpoint.

It is impossible, of course, to give anything approximating a complete account of these subjects in a short article. Thus I strive to give an overview of the structure of these fields by including a few examples from cryptography and developing the theory needed to present the corresponding elliptic curve cryptosystems. Moreover, the paper is intended for one who is familiar with the basic properties of groups and fields, as well as elementary number theory, but with no exposure to cryptography or elliptic curves. Section 2.4 together with Section 2.5.2.2 and Section 2.5.4 sketch the proofs of some key results that follow from the existence of the Weil pairing, an algebraic structure defined on an elliptic curve, and require some additional mathematical maturity. Since Section 2.3 makes no use of Section 2.2, some readers may want to begin with Section 2.3, and use Section 2.2 as a reference when reading Sections 2.4 and 2.5.

2.2 Cryptography

At the heart of cryptography is security—cryptography allows one to very carefully control the information and powers available in a system to various parties in a way that cannot be manipulated or broken by dishonest participants. Thus cryptosystems are often tested for resistance to various "attacks" and required to have a very small probability of being broken. Most of classical cryptography proves results that are conditional on central assumptions. These assumptions are made to be general enough so that they do not rely on the difficulty of a specific problem, such as factoring integers. However, the central assumptions are far from being proven, so cryptography in some sense rests on fragile ground. This section will describe these assumptions as well as some types of cryptosystems that can be constructed under these assumptions. It will also present classical cryptosystems as concrete examples. These cryptosystems will then be modified to use elliptic curves later in this paper. The last subsection will explain some of the mathematics behind the problem of

[†] Shrenik Shah, Harvard '09, is a mathematics concentrator and English minor. He is also enrolled in a concurrent masters program in computer science. He is a founding member of The HCMR and currently serves as Articles Editor.

factoring integers and describe classical algorithms that aim to perform this task more efficiently than the usual brute force methods.

2.2.1 Theory of Computation

The most common formalization for the notion of a "computer" is the Turing machine. For our purposes, we will instead refer nonrigorously to the notion of an **algorithm** and assume that all basic operations, such as addition and multiplication, all take a single time step. These assumptions serve our purposes because we will not analyze the precise complexity of the algorithms we study. The time an algorithm takes is measured as a function of the input size, where the input is a string of binary digits. For the interested reader, [Si] is a good introduction to the formal theory of computation.

An algorithm can either take as an input a string in binary and output "accept" or "reject," or take in a string in binary and output another string in binary. An algorithm of the latter sort is said to be computing a function. Any subset of $\{0,1\}^*$, the set of all binary strings, is termed a **language**. The subset of $\{0,1\}^*$ accepted by an algorithm A is a language said to be **decided** by A. The time $t(n)$ taken by the algorithm is measured as the maximum number of time steps for an input of length n. The class **P** consists of all languages that can be decided by an algorithm that runs in **polynomial time**, meaning that $t(n) \leq p(n)$ for some polynomial p.

Some languages have the property that there exists some string (specific to a particular string $x \in L$) with which membership in the language can be verified quickly. For example, those familiar with basic graph theory will recall the definition of a **Hamiltonian path**, which is a path that visits every vertex exactly once. It is conjectured to be a difficult computational problem to determine, for a given graph G, whether such a path exists. However, given such a path, an algorithm can very easily verify that it is indeed a Hamiltonian path. The description of the path is called a **witness** testifying to the existence of a Hamiltonian path for G.

The class **NP** consists of all languages L with the following property: There exists a polynomial-time algorithm A such that if $x \in L$, then there exists a witness w such that $A(x, w) = 1$. However, if $x \notin L$, then $A(x, w) = 0$ for all choices of w. It is a very well-known open problem to determine whether **P** and **NP** are equal (or not).

In fact, the problem of determining whether a Hamiltonian path exists is called **NP−complete** because it can be proven that for every language L in the class **NP**, there is a polynomial time algorithm that maps any $x \in L$ to graphs that have a Hamiltonian path and $x \notin L$ to graphs that have none. Thus, if a polynomial time algorithm were ever written to determine whether a Hamiltonian path exists, an algorithm could be then developed to decide any language in **NP**. By contrapositive, if any **NP** program were ever proven to have no polynomial time algorithm, then testing for existence of a Hamiltonian path could not be done in polynomial time either.

It seems, then, that proving that $\mathbf{P} \neq \mathbf{NP}$ would be rather powerful, in that it shows that every **NP**-complete problem is difficult to solve. Unfortunately, this difficulty is **worst-case hardness**, which means that for every algorithm, all that is known is that there exists an input on which it is wrong, not that no algorithm can solve the problem for most inputs. Thus, although the idea of basing cryptography on the assumption that $\mathbf{P} \neq \mathbf{NP}$ is an attractive notion, especially since these classes seem "surely" different, a much stronger assumption is needed, as will be seen in the next section.

The last notions needed are merely some technical definitions: A function $\nu : \mathrm{N} \rightarrow \mathbb{R}$ is **negligible** if for any exponent $\alpha \geq 0$ there exists a constant c_α such that $n^{-\alpha} > \nu(n)$ for $n > c_\alpha$. Roughly, if ν eventually shrinks faster than any inverse polynomial, then ν is negligible. We define the **big-O notation** $f(n) = O(g(n))$ to mean that there exists $N \in \mathbf{N}$ and $C > 0$ such that for $n > N$, $|f(n)| \leq C|g(n)|$, and $f(n) = \Theta(g(n))$ to mean that there exists $c, C > 0$ such that $c|g(n)| \leq |f(n)| \leq C|g(n)|$, again for $n > N$.

Finally, a **probabilistic polynomial time** (PPT) algorithm is an algorithm with the additional property that it may generate uniform random bits whenever it wishes. In order words, a PPT algorithm may, in addition to following deterministic instructions, flip a fair two-sided coin.

2.2.2 Central Assumptions

This section details the main assumptions of cryptography. These definitions are taken from [GB]. The main assumption of cryptography is that one-way functions exist. A **one-way function** $f : \{0,1\}^* \to \{0,1\}^*$ has two properties:

- There is a polynomial-time algorithm that computes the one-way function.
- For any PPT A, on a randomly chosen input $x \in \{0,1\}^n$ with uniform distribution, there exists a negligible function ν_A (allowed to depend on A) such that the probability that $A(1^n, f(x)) = z$ where $f(z) = f(x)$ is less than or equal to $\nu_A(n)$ (for n sufficiently large). (Note that $1^n = 1 \ldots 1$ with 1 repeated n times.)

The second property above is a mouthful, so we'll satisfy ourselves with an imprecise definition: given the output on a randomly chosen input of a one-way function, a polynomially bounded algorithm has a low probability of finding an input that would produce the same output. Note that the 1^n in the input to A is to ensure that its running time is based on the size of the space of possible x, rather than on the output $f(x)$, which could be much smaller.

The existence of one-way functions implies that $\mathbf{P} \neq \mathbf{NP}$, but the converse is not true. The field of average-case complexity strives to prove such an equivalence. Unfortunately, many of the results related to this question are pessimistic, and imply that proving an equivalence would require fairly sophisticated, non-intuitive methods.

Some candidate one-way functions:

- Computing the product $n = pq$ of two prime numbers p, q. We will discuss later in this paper some algorithms faster than the naïve $O(\sqrt{n})$ algorithm for factoring such numbers.
- The **discrete logarithm problem**:. Given the multiplicative group $(\mathbb{Z}/p\mathbb{Z})^\times, g^x$, and g, where g is a generator of this group, determine x.
- Computing modular square roots: Given a quadratic residue $a \bmod n$, compute a value x such that $x^2 \equiv a \bmod n$.

As a notational remark, we will use, as above, $a \equiv b \bmod n$ for $n \mid (a-b)$. We will frequently use $=$ for $\equiv$, particularly when working in the ring $\mathbb{Z}/n\mathbb{Z}$. We will also use (a, b) for the greatest common divisor of a and b.

Another problem that is conjectured to be hard is the **Diffie-Hellman problem**. In $(\mathbb{Z}/p\mathbb{Z})^\times$, given the generator g together with g^x and g^y, the problem is to compute g^{xy}. This problem is important in several protocols that we will discuss later. The hardness of this problem is frequently assumed by cryptographers. This is known as the **Diffie-Hellman assumption**. If one has a solution to the discrete logarithm problem for *either* g, g^x or g, g^y, one can certainly compute g^{xy}, so the Diffie-Hellman assumption is *stronger* than the assumption that computing discrete logarithms is hard.

As it turns out, cryptographers do not yet know how to create certain encryption schemes (discussed later) without an even stronger assumption. This is that a **trapdoor function** $f : \{0,1\}^* \to \{0,1\}^*$ exists, with the following properties:

- The function f is one-way.
- There exists a PPT algorithm T so that for every input length n, there exists a $t_n \in \{0,1\}^*$ of length bounded by a polynomial p, and for all $x \in \{0,1\}^n$, $T(f(x), t_n) = z$ with the property that $f(z) = f(x)$.
- It is also important that a trapdoor function can be generated together with its trapdoors t_n efficiently.

Trapdoor functions essentially have a built in method to invert the function, which is important if one party computes the function at a point, while another needs to invert this computation in order to access information. This occurs in **public key encryption**, described in the next section.

2.2.3 Types of Cryptographic Schemes

Cryptographic schemes are simple tasks that represent common tasks requiring security against a very specific breach. This is in contrast to **cryptographic protocols**, which are often rather complex, and composed of many different schemes. Election systems and auction systems are examples of such protocols. In this section, we describe a few of the most common schemes in a nonrigorous fashion, followed by examples. Texts such as [Go1] and [GB] are excellent references for those interested in a formal treatment.

2.2.3.1 Public Key Encryption

In a public key encryption scheme, Alice wants to send a message M to Bob, while an eavesdropper Eve is listening over the channel. Bob generates a secret decryption key d and publishes a public key e. Alice computes the encryption $C = \mathrm{Enc}_e(M)$ and sends it to Bob. Bob receives C and computes $M = \mathrm{Dec}_d(C)$. We require various minimal properties of this encryption system:

- There should be some efficient algorithm to generate pairs (d, e) of a private key and associated public key.
- The algorithms $\mathrm{Enc}_e(\cdot)$ and $\mathrm{Dec}_d(\cdot)$ should be efficient.
- Knowing e and C should not reveal information about M.

In fact, we can define more properties that guard against more subtle attacks. For example, seeing many messages sent across this channel should not reveal additional information about any of the messages. The **one-time pad** discussed in the next section will illustrate the importance of this requirement; it is aptly named, as it should only ever be used once. Another example is **non-malleability**, and requires that an eavesdropper cannot meaningfully modify the message during transmission.

One-time Pad. This does not fit any of the above cryptographic schemes, but is of great historical significance, predating modern cryptography by over half a century. In these earlier times, actual books of bits would be distributed physically, and the sender would indicate which bits were to be used for the decryption. The sheet used would be destroyed after use. The idea is that Alice wants to send a message $M \in \{0,1\}^n$ to Bob, and they have earlier agreed upon a secret one-time pad $s \in \{0,1\}^n$. Alice sends $C = M \oplus s$ to Bob, where $\oplus$ denotes the **exclusive-or** operation, which is defined as bitwise addition modulo 2 of two given binary strings. For example, for any string $t \in \{0,1\}^n$, $t \oplus t = 0^n$. Bob computes $M = C \oplus s = M \oplus s \oplus s$, and both Alice and Bob destroy all traces of s. If Eve knows no information about s, she knows absolutely nothing about M, even if she has C.

The "one-time" property, however, is critical. If Alice sends another message M' with the same key s, then Eve now might have both $C = M \oplus s$ and $C' = M' \oplus s$. By computing $C \oplus C' = M \oplus M'$, Eve gains some information about the messages M and M'. Using some algorithm that perhaps takes advantage of the sparseness of the English language, Eve could possibly decipher M or M' from this information.

RSA Public Key Encryption. In the RSA (Rivest, Shamir, Adleman) cryptosystem, Alice chooses two large primes p, q and releases the number $n = pq$ and an exponent e as her public key, so that $(e, \varphi(n)) = 1$. (We define the **Euler totient function** φ by $\varphi(n) = |\,\{a|(n,a) = 1, 0 \leq a < n\}\,| = |(\mathbb{Z}/n\mathbb{Z})^\times|$.) Then she computes $d = e^{-1} \bmod \varphi(n)$ as her private key. Bob wants to send the message M to Alice, and encrypts it using the function $C = \mathrm{Enc}_{n,e}(M) = M^e \bmod n$. Finally, Alice simply computes $M = \mathrm{Dec}_d(C) = C^d = M^{ed} = M^{k\varphi(n)+1} = M \bmod n$ by Euler's theorem.

Unfortunately, RSA is vulnerable to some attacks. For example, if the same message is sent using e different choices of n, the message can be decrypted via simple Chinese remaindering. This is particularly dangerous when a small e (like 3) is chosen for efficiency purposes. A solution is to pad all messages with random bits at the end, chosen differently with every encryption. Another vulnerability is that if the same message is sent using two relatively prime choices of e for the same value of n, the message can be decrypted using the Euclidean algorithm. A weakness that renders

RSA relatively useless for the purposes of identity-based cryptography is that knowing a pair e, d of encryption and decryption keys allows one to factor n, which we prove in Corollary 5. See [Bo] for an excellent survey of the attacks on this cryptosystem.

Elgamal Public Key Encryption. In Elgamal Public Key encryption, Alice chooses a multiplicative group $\mathbb{F}_p^\times$, a generator g of this group, and a secret integer x. She computes $g_A = g^x$, and publishes as her public key $(\mathbb{F}_p^\times, g, g_A)$. To encrypt $0 < M < p$, Bob sends his message by picking a uniformly random secret integer $0 < y < p - 1$ and computing and sending $C_1 = g^y$, $C_2 = Mg_A^y = Mg^{xy}$ to Alice, who decrypts by computing $C_2C_1^{-x} = Mg^{xy}g^{-xy} = M$.

The key observation is that all of the operations above could be replicated with any cyclic group in place of $\mathbb{F}_p^\times$ above. We will later replace this with a group structure associated to an elliptic curve.

Cryptographically, this system has a key weakness. The pair (C_1, C_2) relate to M by the simple formula $M = C_2C_1^{-x}$. Thus, if Eve has the ability to modify the message as it is sent to Bob, she can change M to the known function of M, $f_k(M) = kM$, for some k, without Bob knowing that any change had been made. Some cryptosystems make it difficult for Eve to control exactly how her modification might affect the sent message—this property is called **non-malleability**. RSA is also malleable, since an encryption of M^2 can be computed from an encryption of M.

Theorem 1. *If the Diffie-Hellman problem is intractible, then Elgamal Public Key Encryption is unbreakable.*

Proof. Suppose for contradiction that given public key g^x, and message (g^y, D) (even for only one such D), one could compute $M = g^{-xy}D$. Then one could compute $M^{-1}D = g^{xy}D^{-1}D = g^{xy}$, violating the intractibility of the Diffie-Hellman problem. □

2.2.3.2 Identity-based Encryption

A weakness of public key encryption is that an adversary can pretend to be Bob, the intended recipient, and request the message to be encrypted in his public key instead of Bob's—he could then decrypt the message received. To solve this problem, one might try to design a trusted database that holds everyone's public keys, but even then, the adversary can intercept communications with the database. The solution provided by **identity-based encryption**, proposed by Shamir in [Sham], is to use some identity-based value that is a function of some unique information about Bob that anyone sending Bob a message would know. As an example, one might use the output of a publicly known hash function (defined in Section 2.2.3.5) on Bob's name, cell phone number, and address as an identifier. A trusted server provides a decryption key to Bob on some occasion in an authenticated manner. Bob can then decrypt messages forever afterwards.

2.2.3.3 Digital Signatures

In a digital signature scheme, Alice wants to sign a document M in a way that is verifiable and unforgeable. She publishes a public key e that anyone can see, and keeps a secret key s. She signs the message with the function $C = \mathrm{Sig}_s(M)$, and sends (M, C) to the desired recipients. Anyone can run the verification algorithm $\mathrm{Ver}_e(M, C)$ to test whether the message was indeed signed by Alice.

There are several levels of information we can assume an adversary might have. In the weakest case, the adversary Eve might know only the public key e, with no examples. More realistically, she might know pairs (M, C) that were produced by Alice earlier. The worst case is that Eve may have forced Alice to sign certain documents of her choice in her efforts to forge a signature on a document that Alice has not yet signed.

We also can place different requirements on possible forgeries, but for the purposes of this paper, we will use the strongest possible requirement, that of **existential unforgeability**: Eve should not be able to sign any message for which she has not yet seen a signature. There are weaker levels of security than this, detailed in [GB].

We finally require, naturally, that $\mathrm{Sig}_s(\cdot)$ and $\mathrm{Ver}_e(\cdot, \cdot)$ should be efficiently computable.

In practice, digital signatures are very frequently attached to emails and other forms of electronic correspondence.

Digital Signature Algorithm. Recall that in a digital signature scheme, Alice wants to equip a document M with an unforgeable signature. The setup for this algorithm is more complex. Alice needs to pick a group $\mathbb{F}_q^\times$ and a large prime p such that $p \mid q-1$, and $\frac{q-1}{p} = e$ is as small as possible. The message will be in the range $0 < M < p$. She also picks an $g \in \mathbb{F}_q^\times$ such that g has order exactly p. By random guesses, exponentiated by a factor of $\frac{q-1}{p}$, this can usually be done quickly. Note that if $q = 2p+1$, meaning that p, q are a pair of Sophie-Germaine primes, then then one can use any nontrivial quadratic residue for g, since the group of residues modulo q is then of order p. Alice also picks a publicly known hash function $H : \{0,1\}^* \to \{0,1\}^\ell$, randomly chooses a secret integer $0 < a < p$, computes $g_A = g^a$, and publishes $(\mathbb{F}_q^\times, p, g, g_A, H)$.

To sign $M \in \{1, \ldots, q-1\}$, Alice computes $H(M)$, picks a random integer $0 < k < p$, and computes $x = g^k \bmod p$ (in $\mathbb{F}_q^\times$ first, then reduced modulo p) and $y = k^{-1}(H(M) + ax) \bmod p$, producing a signed document (M, x, y). The verifier Bob computes $c_1 = y^{-1}H(M) \bmod p$, $c_2 = y^{-1}x \bmod p$, and accepts if $x = g^{c_1}g_A^{c_2} \bmod p$. A correctly signed document is always accepted, since $x = g^k = g^{y^{-1}(H(M)+ax)} = g^{c_1+ac_2} \bmod p$.

This algorithm is frequently used in practice, though it is in fact an open problem to prove that the hardness of breaking this cryptosystem follows from the intractibility of the discrete logarithm problem. On the other hand, the next section describes a provably existentially unforgeable protocol to compute digital signatures.

Elgamal Digital Signatures. Let $q = 2p+1$ be a pair p, q of Sophie-Germaine primes, and G the group of squares in $\mathbb{Z}_q^\times$. Fix a generator g of G. Alice fixes a private key x randomly selected from $\{0, \ldots, p-1\}$, computes $g_A = g^x$, and publishes the public key (p, g, g_A).

To sign $M \in \{1, \ldots, q-1\}$, Alice picks y randomly selected from $\{0, \ldots, p-1\}$, and computes $h = g^y$. Then Alice computes $H(M\|y)$, where $\|$ denotes concatenation, and $c = -xH(M\|h) - y \bmod p$. Finally Alice publishes $\mathrm{SIG}_A(M) = (M, h, c)$.

Verification simply requires checking that $g_A^{H(M\|h)}g^c h = 1$.

Theorem 2. *If computing the discrete logarithm x of g^x is hard, and the hash function is a random oracle, then Elgamal Digital Signatures are existentially unforgeable.*

Proof. If an adversary were to compute (M, h, c) meeting the requirements for a fixed message M, then it would be necessary for the adversary to have obtained $H(M\|h)$ from the oracle, otherwise the value of $g_A^{H(M\|h)}g^c h$ would be a random value. If the adversary knows $H(M\|h)$, this implies that M and h are fixed, since it is difficult (in the random oracle model) to find another message such that $H(M\|h)$ is the same. To forge a single message M, one must be able to determine a value c and a value c' for at least two possible values a, a' for $H(M\|y)$, since $H(M\|y)$ is a random value. Thus, the adversary has $g_A^a g^c h = g_A^{a'} g^{c'} h$, or $g^{x(a-a')} = g^{c-c'}$. The number $a - a'$ has an inverse modulo p since it is nonzero, so $g^x = g^{(c-c')(a-a')^{-1}}$ yields $x \equiv (c-c')(a-a')^{-1} \bmod p$, thus solving the discrete logarithm problem that was assumed hard. □

2.2.3.4 Key Exchange

Diffie and Hellman developed the notion of a **key exchange**, wherein Alice and Bob have no prior shared secret, but they want to be able to send messages via private key cryptography (a cryptographic scheme we will not discuss in this paper). The desired property is that Alice and Bob both compute the same value, and that an eavesdropper Eve is unable to determine anything about that value with high probability. There is a natural generalization to n-partite key exchanges, where n trusted players want to agree on a single key.

This seems fairly straightforward, though there are some interesting issues that arise. With just these properties, Eve could impersonate Alice and share a key with Bob. Bob, thinking that Eve is Alice, will use the shared key to encrypt messages, which Eve can then read. To counter an attack like this, one needs to use authentication, a topic discussed in detail in [GB] and [MvOV].

From the Diffie-Hellman problem described above, it is not difficult to guess their protocol for the key exchange. A group $\mathbb{F}_p^\times$ with generator g is fixed ahead of time. Alice picks $a \in \mathbb{F}_p^\times$

at random, and Bob similarly picks b. Alice sends g^a to Bob, who sends g^b to Alice. They both exponentiate to compute g^{ab}, which is the shared key.

There is an n-partite version of this protocol, again with a publicly known field $\mathbb{F}_p^\times$ with generator g. Players $p_1, \ldots, p_n$ pick secrets $s_1, \ldots, s_n$. In round k, $k = 0, \ldots, n-2$, p_i sends $g^{s_i s_{j_1} \cdots s_{j_k}}$ to all players, where $j_1, \ldots, j_k$ are distinct elements of $\{1, \ldots, i-1, i+1, \ldots, n\}$. The players then use $g^{s_1 \cdots s_n}$ as their shared secret.

2.2.3.5 Cryptographic Hash Functions

Hash functions are a useful tool, often used within the earlier-mentioned cryptosystems. A hash function $H : \{0,1\}^* \to \{0,1\}^*$ sends any string of arbitrary length to its **hash**, of length polynomial in the length of the input (though often of constant length, in practice). In the usual definition of a **collision-free hash function**, it should be intractible to find a **collision** in time polynomial in the input length k, where a collision is a pair of strings s_1, s_2 such that $H(s_1) = H(s_2)$. H is usually required to be a one-way function.

The **random oracle model** of a hash function works as follows: given a message $M \in \{0,1\}^*$, the random oracle always outputs a random string in $\{0,1\}^k$, except when a previous query is repeated, in which case it produces the original output. When using a hash function in a cryptosystem, one sometimes proves results about the security of the cryptosystem by assuming that the hash function is a random oracle. The assumption that a hash function H is indistinguishable from a random oracle is stronger than assuming that H is a collision-free hash function.

2.2.4 Miller-Rabin and Pollard's Algorithm

In a sense that the following theorem makes precise, the group $(\mathbb{Z}/n\mathbb{Z})^\times$ contains all the information about the factorization of n, and this information can be extracted efficiently knowing very little about this group. In fact, simply knowing a reasonably small multiple of its order suffices to factor n in polynomial time.

Theorem 3. *Suppose that we know n and $k\varphi(n)$ for some positive integer k. Assume also that $\log k = \log^{O(1)} n$. Then n can be factored efficiently.*

Proof. It suffices to find a single factor, because given a factorization $n = n_1 \cdot n_2$, one can run the algorithm again on n_1 and n_2, as $\varphi(n_i) \mid \varphi(n)$ (and one can argue that the run time of this recursive algorithm is still polynomial). One can efficiently check whether n is a prime power, so assume this is not the case.

In this case, we can design an algorithm as follows. Given a composite input n, we first factor $k\varphi(n) = 2^\ell m$ where $\ell \in \mathbb{Z}$ is chosen maximally such that $m \in \mathbb{Z}$. Next, we pick a random integer a, where $1 < a < n$, and compute (a, n). If $(a, n) \neq 1$ then we are done, since in this case we have found a nontrivial factor of n. Otherwise, $(a, n) = 1$ and

$$a^{\varphi(n)} \equiv a^{k\varphi(n)} \equiv a^{2^\ell m} \equiv 1 \bmod n.$$

Thus, if we consider

$$a^m, a^{2m}, a^{2^2 m}, \ldots, a^{2^\ell m} \bmod n$$

we have a sequence that eventually becomes 1. For fixed a, consider the largest value of j such that $a^{2^j m} \not\equiv 1 \bmod n$. In Lemma 4, we show that for at least $\frac{1}{2}$ of the possible choices of a, we have that $a^{2^j m} \not\equiv -1 \bmod n$. Then, $a^{2^{j+1} m} - 1 \equiv (a^{2^j m} + 1)(a^{2^j m} - 1) \equiv 0 \bmod n$. Thus, $(a^{2^j m} + 1)(a^{2^j m} - 1) = sn$ for some integer s. Also, neither of the factors on the left are 0 or multiples of n, since we required $a^{2^j m} \not\equiv -1 \bmod n$. So we can compute $(a^{2^j m} + 1, n)$ and $(a^{2^j m} - 1, n)$ to obtain a factorization of n into $n_1 \cdot n_2$, as desired. By choosing random values for a until we find one that yields a factorization of n, this algorithm terminates in expected polynomial time. □

Lemma 4. *In the proof of Theorem 3 above, at least $\frac{1}{2}$ of our choices of a have the property that there exists j such that $a^{2^j m} \not\equiv 1, -1 \bmod n$ while $a^{2^{j+1} m} \equiv 1 \bmod n$. Moreover, we will use only the fact that $a^{2^\ell m} \equiv 1 \bmod n$ for all $a \in (\mathbb{Z}/n\mathbb{Z})^\times$.*

Proof. We adapt the proof of correctness of the Miller-Rabin primality test from [CLRS]. Define a pair (a, j) of integers to be **bad** if $a \in (\mathbb{Z}/n\mathbb{Z})^\times, j \in \{0, 1, \ldots, \ell\}$, and $a^{2^j m} \equiv -1 \bmod n$. The desired result translates into proving that the number of bad pairs (a, j) is at most $\frac{1}{2}|(\mathbb{Z}/n\mathbb{Z})^\times|$, since for any a of the form described in the statement of the lemma, there is a unique value of j such that (a, j) is bad. Note that since m is odd, $(n-1, 0)$ is bad. Thus there exists at least one bad pair, so we may pick the largest possible j such that there is a bad pair (a, j), and fix a value of a so that (a, j) is bad. Note that $j < \ell$ since, by assumption, $a^{2^\ell m} \equiv 1 \bmod n$ for all a. Let

$$S = \left\{ x \in (\mathbb{Z}/n\mathbb{Z})^\times | x^{2^j m} \equiv \pm 1 \bmod n \right\}.$$

This set is closed under multiplication modulo n, so it is a subgroup of $(\mathbb{Z}/n\mathbb{Z})^\times$. Every bad pair (a, j) has a a member of S, because we picked a to be maximal and we allow $x^{2^j m} \equiv \pm 1 \bmod n$ in the definition of S. If S also contained numbers a such that (a, j) is not bad, this would be fine, as we are only proving a bound on the number of bad pairs.

We now prove that $S \neq (\mathbb{Z}/n\mathbb{Z})^\times$. Note that n by assumption is not a prime power, so $n = n' \cdot p^\alpha$ for some prime $p \mid n$, where α is chosen maximally with $p^\alpha \mid n$, and $n' \neq 1$. Since $a^{2^j m} \equiv -1 \bmod n$, we have $a^{2^j m} \equiv -1 \bmod n'$ by the Chinese Remainder Theorem, as $(n', p^\alpha) = 1$. Moreover, again by this theorem, there exists b such that $b \equiv a \bmod n', b \equiv 1 \bmod p^\alpha$. Thus, by our preceding calculation, $b^{2^j m} \equiv -1 \bmod n', b^{2^j m} \equiv 1 \bmod p^\alpha$. By the Chinese Remainder Theorem, $b^{2^j m} \not\equiv 1 \bmod n'$ implies $b^{2^j m} \not\equiv 1 \bmod n$, and $b^{2^j m} \not\equiv -1 \bmod p^\alpha$ implies $b^{2^j m} \not\equiv -1 \bmod n$. Thus $b^{2^j m} \not\equiv \pm 1 \bmod n$, so $b \notin S$.

It suffices, then, to show that $b \in (\mathbb{Z}/n\mathbb{Z})^\times$. Note that since $a \in (\mathbb{Z}/n\mathbb{Z})^\times$, $(a, n) = 1$, so $(a, n') = 1$. Since $b \equiv a \bmod n'$, $(b, n') = 1$. Also by definition, $b \equiv 1 \bmod p^\alpha$, so $(b, p^\alpha) = 1$. Thus, b is relatively prime to both n' and p^α, so b is relatively prime to their product, n. Thus $b \in (\mathbb{Z}/n\mathbb{Z})^\times$, as desired. So S is strictly contained in $(\mathbb{Z}/n\mathbb{Z})^\times$, and by Lagrange's theorem, it has order at most $\frac{1}{2}|(\mathbb{Z}/n\mathbb{Z})^\times|$. □

Corollary 5. *If we have a pair e, d of corresponding RSA encryption and decryption keys, then we can factor n.*

Proof. Since $ed \equiv 1 \bmod \varphi(n)$, $ed - 1$ is a multiple of $\phi(n)$, whereby we can factor n using Theorem 3. □

A **Carmichael number** is a number n such that $a^{n-1} \equiv 1 \bmod n$ for all integers a such that $(a, n) = 1$.

Corollary 6. *If we have a bound of $\log^{O(1)} n$ on the size of the primes dividing $\varphi(n)$, or if n is a Carmichael number, we can factor n efficiently.*

Proof. The first statement can be shown by using a weak estimate on the density of primes to find an upper bound on the product of primes smaller than the bound $\log^{O(1)} n$ in order to find a small multiple of $\varphi(n)$. The second follows from the theorem that for a prime p dividing a Carmichael number n, $p - 1 \mid n - 1$. □

We can similarly define the Rabin-Miller randomized polynomial-time primality test: Given a positive integer n, we first check that it is not a perfect power. We then factor $n - 1 = 2^\ell m$, where ℓ is maximal, pick a random $1 < a < n$, and compute (a, n). If $(a, n) \neq 1$ then we're done, since n is then composite. Otherwise, $(a, n) = 1$. We then check that $a^{2^\ell m} \equiv a^{n-1} \equiv 1 \bmod n$, which can fail only if n is composite. We then consider

$$a^m, a^{2m}, a^{2^2 m}, \ldots, a^{2^\ell m} \bmod n,$$

a sequence that eventually becomes 1. If $a^{2^j m} \neq \pm 1$ while $a^{2^{j+1}m} = 1$, then $x^2 - 1$ has more than two roots in $\mathbb{Z}/n\mathbb{Z}$, so this ring cannot be a field and n cannot be prime.

Corollary 7. *If this test outputs composite, it is correct. If n is prime, the test outputs prime with probability $\geq \frac{1}{2}$.*

Proof. The first sentence is clear from the algorithm. So suppose n is composite. Note that the $a \in (\mathbb{Z}/n\mathbb{Z})^\times$ with $a^{n-1} \equiv 1 \bmod n$ form a subgroup T of $(\mathbb{Z}/n\mathbb{Z})^\times$. We divide into cases, depending on whether or not $T = (\mathbb{Z}/n\mathbb{Z})^\times$.

If there exists any $x \in (\mathbb{Z}/n\mathbb{Z})^\times \setminus T$, then by Lagrange's theorem, at least $\frac{1}{2}$ of the choices of x will have this property, since the smallest possible index for a proper subgroup is 2. If our algorithm as described above picks any $a \in (\mathbb{Z}/n\mathbb{Z})^\times \setminus T$, which it does with probability $\geq \frac{1}{2}$, it correctly classifies n.

If, instead, $T = (\mathbb{Z}/n\mathbb{Z})^\times$, $n-1$ has all of the necessary properties that $k\varphi(n)$ had in the proof of Lemma 4 (as remarked in its statement). This lemma then shows that at least $\frac{1}{2}$ of the choices of a will satisfy $a^{2^j m} \neq -1 \bmod n$ for the largest value of j such that $a^{2^j m} \neq 1 \bmod n$, and will thus lead the algorithm to classify n as composite. □

We can exploit the fact above by essentially "guessing" the structure of $(\mathbb{Z}/n\mathbb{Z})^\times$, trying to pick a multiple of $\varphi(n)$ by choosing numbers with many small prime divisors. The homomorphism $\mathbb{Z}/n\mathbb{Z} \rightarrow \mathbb{Z}/p_i\mathbb{Z}$ given by sending a to its residue modulo p implies that raising a number $a \in (\mathbb{Z}/n\mathbb{Z})^\times$ such that $p \nmid a$ to a power that is a multiple of $p_i - 1$ will yield a multiple of p_i.

This leads to a Pollard's factorization algorithm: We pick an integer $0 < a < n$, pick an integer k that is a multiple of many small primes (say $\operatorname{lcm}(1, \ldots, m)$), and compute $(a^k - 1, n)$. If $(a^k - 1, n) = n$, we use the algorithm above to factor n with k in place of $k\varphi(n)$, and if $(a^k - 1, n) = 1$, we pick a larger value of k (or a larger choice of m). If $p_i - 1 \mid k$ for some but not all i, $(a^k - 1, n)$ is likely to be a proper nontrivial factor of n, as $a^k - 1$ is unlikely to be a multiple of $p_j - 1$, where $p_j - 1 \nmid k$ (it is not difficult to see that this occurs with probability $\Theta(\frac{1}{p_j - 1})$). If $(a^k - 1, n) = 1$, then increasing k is natural, because this could only occur if k were not a multiple of $p_i - 1$ for any prime p_i.

Unfortunately, this algorithm sometimes will be very inefficient, particularly when the $p_i - 1$ are not products of small primes. This observation illustrates the failure of $(\mathbb{Z}/n\mathbb{Z})^\times$ to reveal the factorization of n easily, even though its order alone would be sufficient to factor n. Although in most cases, the order $\varphi(n)$ is a product of mostly small primes, in the worst case, when $\varphi(n)$ is a multiple of some very large primes, this approach gives us little hope. Thus Pollard's algorithm is an efficient way to factor most n, but not all.

We thus see that the general problem of determining the properties of $(\mathbb{Z}/n\mathbb{Z})^\times$ is one that may be easy for most randomly chosen inputs, but very hard on certain inputs. It would be nice, then, to have a group that contains all of the information about the factorization of n, yet whose order is "random." If this were the case, there would be a good chance of being able to factor n by trying many times on randomly chosen orders, using the same methodology as Pollard's algorithm.

This is where elliptic curves come in handy. Each elliptic curve E over the ring $\mathbb{Z}/n\mathbb{Z}$ associates to n a group $G(E, \mathbb{F}_{p_i})$ that, in a sense made precise by Theorem 8, contains the same information about n as the group $(\mathbb{Z}/n\mathbb{Z})^\times$. Moreover, we will see that the role of $p_i - 1$ in this problem is replaced by a number in the interval $p_i + 1 - 2\sqrt{p_i} < |G(E, \mathbb{F}_{p_i})| < p_i + 1 + 2\sqrt{p_i}$. It is rather likely that some number in this range will have small prime factors, and as a consequence, the algorithm will much more efficiently be able to factor numbers.

2.3 Elliptic Curves

Elliptic curves may be viewed from the perspectives of many fields of mathematics, including number theory, analysis, and algebraic geometry. Although we'll give a taste of this in Section 2.4, the focus of this section will be on the special case of elliptic curves over a finite field. Our goal is to define the **group law** associated to an elliptic curve over a field and remark upon the generalization to a ring such as $\mathbb{Z}/n\mathbb{Z}$.

2.3.1 Preliminaries

We first define affine and projective space over a field K. The **affine space** $\mathbf{A}^n_K$ is defined to be the set K^n, with no vector space structure. The **projective space** $\mathbf{P}^n_K$ is defined, again as a set, to be $\{(\alpha_0, \ldots, \alpha_n) \in K^{n+1} \setminus 0\}$ modulo the equivalence relation $(\alpha_0, \ldots, \alpha_n) \sim (\lambda\alpha_0, \ldots, \lambda\alpha_n)$ for $\lambda \in K^\times$. We will write the equivalence class of $(\alpha_0, \ldots, \alpha_n) \in \mathbf{P}^n_K$ as $[\alpha_0, \ldots, \alpha_n]$. Intuitively, the projective space contains points "at infinity" corresponding to intersections of parallel hyperplanes in affine space. On the other hand, it is clear from the definition that there are no distinguished points in projective space, and so there is no preferred affine subset of projective space. A cover of projective space by subsets "isomorphic" to affine space can be found by considering the $n+1$ subsets where $\alpha_i = 0$ for each i. These are called $\mathbf{A}^n_{K,i}$, though the K is often omitted. The technicalities of defining the structure on these spaces, which are actually **varieties**, can be found in [Shaf] or [Ha]. None of these details will be important to the overview in this paper, though it is occasionally useful to keep in mind that the endomorphisms on elliptic curves we consider in 2.4 are cases of a more general notion.

An elliptic curve is defined over a field or ring. Let the field K have characteristic neither equal to 2 or 3, an assumption we will hold throughout this paper. Then an **elliptic curve** E is the set of points $[x, y, z] \in \mathbf{P}^2_K$ defined by the homogeneous polynomial $y^2z = x^3 + axz^2 + bz^3$ over the projective plane $\mathbf{P}^2_K$ for $a, b \in K$, or the curve $y^2 = x^3 + ax + b$ defined over the affine plane $\mathbf{A}^2_K$ together with a point at infinity, corresponding to $[0, 1, 0]$ in the projective form of the curve. One should also think of the defining polynomial itself as being part of the information present in E. Under the conditions on the characteristic, any other object one might call an "elliptic curve" can be transformed by a simple change of variables into the form given. Over a ring R, the definition is similar, though in this case, we will require $2, 3 \in R^\times$. The projective space $\mathbf{P}^2_R$ will also need to be redefined, which we do in Section 2.3.3. Note that we say "elliptic curve over K" (or R) to mean an elliptic curve whose coefficients a, b are in K (or R).

2.3.2 The Group Law

Suppose that we have points $P_1 = (x_1, y_1), P_2 = (x_2, y_2)$ on an elliptic curve $y^2 = x^3 + ax + b$ with $a, b \in \mathbb{Q}$ (or, more generally, any field K). Then an equation for the line through P_1 and P_2 is obtained by taking $m = \frac{y_1 - y_2}{x_1 - x_2}$ to be the slope of this line and using the formula $y = m(x - x_1) + y_1$. This line intersects the curve in a third point, which we shall solve for by substituting this expression for y: $(m(x - x_1) + y_1)^2 = x^3 + ax + b$, or

$$x^3 - m^2x^2 + (-2my_1 + 2m^2x_1 + a)x - m^2x_1^2 + 2mx_1y_1 - y_1^2 + b = 0.$$

Since the coefficient m^2 is the sum of the roots of the polynomial, while x_1, x_2 are already roots, we can find the abscissa of the third point as $x_3 = m^2 - x_1 - x_2$, and the ordinate from the equation of the line, $y_3 = m(x_3 - x_1) + y_1$. We label $P_3 = (x_3, y_3)$, and we define $0 = P_1 + P_2 + P_3$ in the additive group of points on the elliptic curve, so that $P_1 + P_2 = -P_3$. We define the negation of a point to be its reflection over the x-axis, meaning that $-(x, y) = (x, -y)$. The third point on this line connecting (x, y) and $(-x, y)$ is the point at infinity, and the identity of this additive group.

Commutativity and identity properties follow immediately from the above definition, but associativity is tedious to verify, as it breaks up into cases. We refer a reader interested in the proof to [ST], though we will provide a somewhat opaque proof of this result for subfields of $\mathbb{C}$ when we motivate the Weil pairing in Section 2.4.

Associated to an elliptic curve E over a field K we have defined an abelian group structure $G(E, K)$ on its points. Moreover, for a field extension L/K, there is a natural injection $G(E, K) \hookrightarrow G(E, L)$. Many amazing results in number theory describe various properties of $G(E, \mathbb{Q})$. Although we will not need them, they are worth mentioning. Mordell's theorem shows that this group is always finitely generated. Mazur's theorem shows that the torsion subgroup is either C_n for $1 \leq n \leq 10$ or $n = 12$, or $C_2 \times C_{2n}$ for $1 \leq n \leq 4$. Determining $G(E, \mathbb{F}_p)$ is generally easier than computing the group law over $\mathbb{Q}$, and in fact, the most important information about $G(E, \mathbb{F}_p)$ for our purposes will be its order (which is finite).

2.3.3 Elliptic Curves over $\mathbb{Z}/n\mathbb{Z}$

One can define elliptic curves over rings as well, though to do so requires a more complicated group law, since division is not permissible and other issues arise. The story is detailed in Lenstra's paper [Le], which also describes the factoring algorithm we present in Section 2.5.1. There are various constraints that make the process simpler. We'll avoid delving into the general theory and study the specific case of $\mathbb{Z}/n\mathbb{Z}$, where n is prime to 2 and 3. This will suffice for our purposes, because our factoring algorithm will specifically check for divisibility by small primes. Over a ring R, the projective space $\mathbf{P}^n_R$ is redefined as

$$\left\{(\alpha_0,\dots,\alpha_n)\in R^{n+1}\middle|(\alpha_0,\dots,\alpha_n)=(1)\right\}/(\alpha_0,\dots,\alpha_n)\sim(\lambda\alpha_0,\dots,\lambda\alpha_n),\lambda\in R^\times,$$

and it becomes generally more important to consider the elliptic curve within projective space. We still define the curve as the solutions to the homogeneous equation

$$y^2z=x^3+axz^2+bz^3$$

in $\mathbf{P}^2_R$, though we require that the discriminant $-4a^3-27b^2$ be a unit (which just means nonzero over a field, so this is consistent).

The ring $\mathbb{Z}/n\mathbb{Z}$ holds information about its factors within its subgroups, and has the property that if $n=n_1n_2$, where $(n_1,n_2)=(1)$, then $\mathbb{Z}/n\mathbb{Z}=(\mathbb{Z}/n_1\mathbb{Z})\times(\mathbb{Z}/n_2\mathbb{Z})$. We can show that this translates to the elliptic curve as well. Note that for an elliptic curve E over R and an injection $\varphi: R\hookrightarrow S$, $G(E,R)\subseteq G(E,S)$. But we can replace the injection φ with any homomorphism, and apply this map both to the coefficients of E and to R to obtain a homomorphism $\tilde{\varphi}:G(E,R)\to G(E,S)$. In particular, for an elliptic curve E over $\mathbb{Z}/n\mathbb{Z}$ and φ_i reduction modulo n_i, we obtain maps $\tilde{\varphi}_i:G(E,\mathbb{Z}/n\mathbb{Z})\to G(E,\mathbb{Z}/n_i\mathbb{Z})$. We can now prove:

Theorem 8 ([Wa, pp. 65–66]). *The map*

$$\tilde{\varphi}_1\times\tilde{\varphi}_2:G(E,\mathbb{Z}/n\mathbb{Z})\to G(E,\mathbb{Z}/n_1\mathbb{Z})\times G(E,\mathbb{Z}/n_2\mathbb{Z})$$

is an isomorphism.

Proof. Note first of all that the discriminant of E when reduced modulo n_1 or n_2 is still a unit, since it is relatively prime to n. Thus the groups on the right above are well-defined. Reduction modulo n_i yields, via the Chinese Remainder Theorem,

$$\varphi_1\times\varphi_2:\mathbb{Z}/n\mathbb{Z}\cong\mathbb{Z}/n_1\mathbb{Z}\times\mathbb{Z}/n_2\mathbb{Z},$$

and thus a bijection between triples of these elements. This map induces maps on $\mathbf{P}^2_{\mathbb{Z}/n\mathbb{Z}}$, which we'll denote $\overline{\varphi}_1,\overline{\varphi}_2$, since if $(x,y,z)=(ux',uy',uz')$ are equivalent triples, their images are equivalent via the image of the unit u. When applied to the bijection on triples, this yields a bijection

$$\overline{\varphi}_1\times\overline{\varphi}_2:\mathbf{P}^2_{\mathbb{Z}/n\mathbb{Z}}\to\mathbf{P}^2_{\mathbb{Z}/n_1\mathbb{Z}}\times\mathbf{P}^2_{\mathbb{Z}/n_2\mathbb{Z}}.$$

Finally, $y^2z=x^3+axz^2+bz^3 \bmod n$ implies $y^2z=x^3+axz^2+bz^3 \bmod n_i$ for $i=1,2$. The converse is true, again by the Chinese Remainder Theorem. Thus the map $\tilde{\varphi}_1\times\tilde{\varphi}_2$ is a bijection. That it is a homomorphism can be verified by a tedious but simple computation that we will omit. □

Our last comment is that although we have not specified the exact group law over a ring $\mathbb{Z}/n\mathbb{Z}$, one can use the usual group law over a field K when dealing only with units. This fails to yield the right answer exactly when a nonzero nonunit appears in one of the coordinates of one of the points. For the purpose of factoring, however, we actually need not explain how to carry out the group operation in this situation, as a nonzero nonunit appears exactly when we have successfully managed to factor n, so the algorithm can terminate at this stage.

2.4 The Weil Pairing

The Weil pairing is a construction on the n-torsion points of an elliptic curve that is important for proving theoretical results, such as the analogue of the Riemann Hypothesis for elliptic curves, as well as for computational applications, discussed in Section 2.5.2.2 and Section 2.5.4. The progression used in the subsections following the motivation come largely from Chapters 2 to 4 and 11 to 12 of [Wa], though we provide few proofs. Our goal is to sketch the ideas behind the results on elliptic curves over finite fields most relevant to cryptography. Readers without background in complex analysis can safely skip Section 2.4.1.

The texts [Si1], [Si2] provide an abstract viewpoint on pairings that is more useful from the purposes of number theory, while [CFA] provides a comprehensive treatment of the cryptographic applications of the Weil pairing and the related Tate-Lichtenbaum pairing.

2.4.1 Motivation

For those familiar with either the theory of algebraic curves over $\mathbb{C}$ or of compact Riemann surfaces, the following will serve as motivation for some of the theorems that follow (and provide proofs of special cases of results we will not prove).

If we view an elliptic curve in $\mathbf{P}^2_{\mathbb{C}}$ as a compact Riemann surface X of genus 1, there exists a lattice $\Lambda \subseteq \mathbb{C}$ such that X maps biholomorphically to the quotient $E = \mathbb{C}/\Lambda$. Indeed, for the inverse, the Weierstrass $\wp$-function defines the map $z \mapsto (\wp(z), \wp'(z))$ sending $\mathbb{C}/\Lambda$ to the solutions of $y^2 = 4x^3 - 60G_4(\Lambda)x - 140G_6(\Lambda)$, where G_4, G_6 are the Eisenstein invariants of the lattice Λ.

The importance of this perspective comes from the obvious addition law on points in $\mathbb{C}/\Lambda$, which is carried via the isomorphism to yield an abelian group structure on E. This formula is given by

$$\wp(z_1 + z_2) = \frac{1}{4}\left(\frac{\wp'(z_2) - \wp'(z_1)}{\wp(z_2) - \wp(z_1)}\right)^2 - \wp(z_1) - \wp(z_2).$$

One can derive the associativity of the group law over subfields of $\mathbb{C}$ from this isomorphism. We also can easily determine the n-torsion points of $\mathbb{C}/\Lambda$: if $\Lambda = \mathbb{Z}\omega_1 \oplus \mathbb{Z}\omega_2$, then these are precisely $\frac{\ell_1\omega_1}{n} + \frac{\ell_2\omega_2}{n}$, for $0 \leq \ell_1, \ell_2 < n$. Thus the n-torsion group, which we denote by $G_n(E, \mathbb{C})$, is isomorphic to $\mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$, for any n. More generally, we let $G_n(E, K)$ denote the subgroup of $G(E, \overline{K})$ (where $\overline{K}$ denotes the algebraic closure) of points P such that $nP = 0$. With some additional technical steps, one can use the case of $\mathbb{C}$ to conclude that $G_n(E, K) \cong \mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$ for general fields of characteristic 0.

We denote by $C_n = \langle\zeta\rangle$ the group of n^{th} roots of unity in $\mathbb{C}$, where ζ is a primitive root. In [Ga], a Weil pairing $w_n : G_n(E, \mathbb{C}) \times G_n(E, \mathbb{C}) \to C_n$ is defined by the formula:

$$w_n\left(\frac{\ell_1\omega_1}{n} + \frac{\ell_2\omega_2}{n}, \frac{m_1\omega_1}{n} + \frac{m_2\omega_2}{n}\right) = \exp\left(\frac{2\pi i(\ell_1 m_2 - m_1\ell_2)}{n}\right)$$

The key properties of this pairing follow directly from this definition. It is linear in each component, and satisfies $w_n(0, g) = w_n(g, 0) = w_n(g, g) = 1$ for all g. On the other hand, for $g \neq 0$, there exist h, h' such that $w_n(g, h) \neq 0$, $w_n(h, g) \neq 0$—this means that w_n is **nondegenerate**. Finally $w_n(g, h) = w_n(h, g)^{-1}$. In the terminology of linear algebra, one will easily recognize that these properties are analogous to those of a nondegenerate skew-symmetric bilinear form.

2.4.2 Endomorphisms

We next study certain maps on elliptic curves. These maps are a general case of **regular maps**. Let us consider a curve in affine space, $\mathbf{A}^2_K$. Then the coordinates x, y themselves define functions assigning to each point of the curve a value in K. One can generate many functions on the curve by considering polynomials in x, y. Such maps are examples of **regular functions**. One can also define a map $\tau : E \to \mathbf{A}^2_K$ by defining $\tau(x, y) = (p_1(x, y), p_2(x, y))$ for a pair of regular functions p_1, p_2. If the image of an elliptic curve E under a map of this form happens to lie on E, and the map acts as a group homomorphism on $G(E, K)$, then the map is called a **endomorphism**.

We allow, in fact, p_1 and p_2 to be rational functions. Under a map of this form, it is possible for points of the curve to be sent to the point at infinity.

We can determine a convenient form in which all such maps can be written. Since all points on the curve E satisfy $y^2 = x^3 + ax + b$, we can assume that the numerator and denominator of p_1, p_2 have no powers of y larger than 1. In fact, multiplying the denominator by its conjugate with respect to y (meaning we replace y with $-y$), we find that $p_i(x,y) = \frac{q_i(x)+s_i(x)y}{t_i(x)}$ for functions $q_i, s_i, t_i \in K(x,y)$ and $i \in \{1,2\}$. Since $\tau(x,-y) = \tau(-(x,y)) = -\tau(x,y)$, in fact, we have $s_1(x) = 0$ and $q_2(x) = 0$, so $\tau(x,y) = (r_1(x), r_2(x)y)$, where $r_1, r_2 \in K(x)$ are rational functions of x.

If the map τ is nonzero, we define the **degree** of this map to be the larger among the degree of the numerator and denominator of r_1, and if τ is zero, we define the degree to be 0. The sum of endomorphisms defines an endomorphism by summing the images using the group law, and endomorphisms admit multiplication by an integer in the same way. We denote the identity endomorphism $\tau(x,y) = (x,y)$ by 1 and, over $\mathbb{F}_q$, the **Frobenius endomorphism** by $\varphi_q(x,y) = (x^q, y^q)$, easily checked to be an endomorphism.

Following [Wa], we will use the Weil pairing in the next section to derive the following result:

Theorem 9. *Let σ, τ be endomorphisms and $s, t \in \mathbb{Z}$. Then*

$$\deg(s\sigma + t\tau) = s^2 \deg \sigma + t^2 \deg \tau + st(\deg(\sigma + \tau) - \deg \sigma - \deg \tau).$$

Another way to create new endomorphisms is via **composition**, for example $\varphi_q^n = \varphi_q \circ \cdots \circ \varphi_q$, n times. The endomorphism τ is defined to be **separable** if $r_1'(x)$ is not identically 0. This rather technical definition is explained by Lemma 10, which expresses the notion of separability in terms of the degree and kernel of the map, two seemingly more intrinsic qualities. This lemma can be proven by carefully keeping track of the roots of the numerator and denominator of r_1, r_2.

Lemma 10. *([Wa], p. 49-50) For τ a nonzero endomorphism on an elliptic curve over an algebraically closed field, $\deg \tau \geq |\ker \tau|$, with equality if and only if τ is separable.*

Thus, if we wanted to compute the number of points on the curve E over $\mathbb{F}_{p^k}$, we might consider φ_p^k, which fixes $\mathbb{F}_{p^k}$ and no other elements of $\overline{\mathbb{F}}_p$. Then the endomorphism $\varphi_{p^k} - 1$ has kernel exactly corresponding to the points of E over $\mathbb{F}_{p^k}$. However, we need to know whether this endomorphism is separable or not. This question is answered by another lemma, a convenient condition for separability of certain endomorphisms over $\mathbb{F}_{p^k}$.

Lemma 11. *([Wa], p. 54) Over $\mathbb{F}_{p^k}$, where $r, s \in \mathbb{Z}$, $r\varphi_{p^k} + s$ is separable exactly when $p \nmid s$.*

The main idea of the proof is to first reduce to the question of whether s is separable, and to then answer this by first showing that if $n = (r_1(x), r_2(x)y)$, $\frac{r_1'(x)}{r_2(x)} = n$.

These results will be used later to conclude the Hasse bound. The final step simply involves computing the degree of $\varphi_{p^k} - 1$ using Theorem 9.

Another result that we will invoke later is:

Theorem 12 ([Wa, p. 95–96]). *If $|G(E, \mathbb{F}_{p^k})| = p^k + 1 - \ell$, then $\varphi_{p^k}^2 - \ell\varphi_{p^k} + p^k = 0$ as endomorphisms. Moreover, for $m \neq \ell$, $\varphi_{p^k}^2 - m\varphi_{p^k} + p^k \neq 0$.*

For the first statement, the key idea is to show that $\varphi_{p^k}^2 - \ell\varphi_{p^k} + p^k$ is identically zero on $G_n(E, \mathbb{F}_p)$ for infinitely many choices of n. Then Lemma 10 implies that this endomorphism is identically zero. The second is easy: If $\varphi_{p^k}^2 - m\varphi_{p^k} + p^k = 0$, then by subtracting $\varphi_{p^k}^2 - \ell\varphi_{p^k} + p^k = 0$ we find $(\ell - m)\varphi_{p^k} = 0$ as endomorphisms, which can happen only if $\ell = m$.

2.4.3 The Weil Pairing and its Consequences

We shall state two key theorems, whose proofs are in [Wa].

Theorem 13 ([Wa, p. 75]). *If $\operatorname{char} K \nmid n$ or $n = 0$, then $G_n(E, K) \cong \mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$.*

Theorem 14 ([Wa, pp. 83, 334-335]). *Let $C_n \subset \overline{K}$ be the group of n^{th} roots of unity and E be the elliptic curve $y^2 = x^3 + ax + b$. Then there exists a map $w_n : G_n(E, K) \times G_n(E, K) \to C_n$ called the **Weil pairing**, with the properties that*

- *The map w_n is bilinear and nondegenerate in each variable (as defined in Section 2.4.1).*
- *We have $w_n(g, g) = 1$ and $w_n(g, h) = w_n(h, g)^{-1}$ for $g, h \in G_n(E, K)$.*
- *If $\varphi : \overline{K} \to \overline{K}$ is an automorphism with a, b as fixed points, $w_n(\varphi(g), \varphi(h)) = \varphi(w_n(g, h))$.*
- *If σ is an endomorphism of E, $w_n(\sigma(g), \sigma(h)) = w_n(g, h)^{\deg \sigma}$.*

If σ is an endomorphism of E, it is easy to see that it restricts to an endomorphism of $G_n(E, K)$. Using Theorem 13 we can choose a basis ω_1, ω_2 for $G_n(E, K)$, so that σ is defined by $(\omega_1, \omega_2) \mapsto (\ell_1\omega_1 + \ell_2\omega_2, m_1\omega_1 + m_2\omega_2)$. Then one can show via simple computations (see [Wa]) that $w_n(\omega_1, \omega_2)$ is a primitive n^{th} root of unity and that $\det \left(\begin{smallmatrix} \ell_1 & \ell_2 \\ m_1 & m_2 \end{smallmatrix}\right) = \deg \alpha$. As a consequence, one can prove that Theorem 9 holds over fields of characteristic p by computing the determinants of both sides as endomorphisms over $G_n(E, K)$ for each n with $p \nmid n$. One can extend this result to all n by applying Lemma 10.

2.4.4 The Hasse Bound

Over finite fields, it is possible to very accurately determine the order of the group of points on an elliptic curve. We will use some of the facts about endomorphisms we stated without proof in Section 2.4.2.

Let E have coefficients in a finite field $\mathbb{F}_{p^k}$. By the remarks above, which used Lemma 10 and 11, $\deg(\varphi_{p^k} - 1) = |G(E, \mathbb{F}_{p^k})|$.

Theorem 15 ([Wa, pp. 91–94]). *For an elliptic curve E over $\mathbb{F}_{p^k}$, $2\sqrt{p^k} \geq p^k + 1 - |G(E, \mathbb{F}_{p^k})| \geq -2\sqrt{p^k}$.*

Proof. Define $q = p^k$ for simplicity. By the preceding observation, $q + 1 - |G(E, \mathbb{F}_q)| = q + 1 - \deg(\varphi_q - 1)$, a quantity we will call ℓ. By Theorem 9, for $s, t \in \mathbb{Z}$ with $p \nmid t$,

$$\begin{aligned} \deg(s\varphi_q - t) &= s^2 \deg(\varphi_q) + t^2 \deg(-1) + st(\deg(\varphi_q - 1) - \deg(\varphi_q) - \deg(-1)) \\ &= qs^2 + t^2 - st\ell \end{aligned}$$

Since $\deg(s\varphi_q - t) = qs^2 + t^2 - st\ell \geq 0$, $q\left(\frac{s}{t}\right)^2 - \ell\left(\frac{s}{t}\right) + 1 \geq 0$. Since $\left\{\frac{s}{t} : p \nmid t\right\}$ is dense in $\mathbb{R}$, this inequality holds for all real $r \in \mathbb{R}$ in place of $\frac{s}{t}$, so $qr^2 - \ell r + 1 \geq 0$. This implies that the discriminant is nonpositive, so $\ell^2 - 4q \leq 0$, or $|\ell| \leq 2\sqrt{q}$, as desired. □

This powerful result has many important cryptographic consequences. Several algorithms use the narrow range of possible orders for $G(E, \mathbb{F}_q)$ given by the Hasse bound in an essential way to compute the order of the group. We also can use the range for the order to bound the running time of algorithms, such as those for computing factorizations and discrete logarithms.

2.4.5 The Riemann Hypothesis for Elliptic Curves

Note that the following results are entirely optional, as the they are not used in the remainder of the paper. The purpose of this section is to prove that the Hasse bound implies an analogue of the Riemann hypothesis for curves. This connection is suggestive of a far more general analogy between points on varieties and primes in number fields. The usual Riemann hypothesis is equivalent to a statement about the density of primes. In the same sense, the Riemann hypothesis on an elliptic curve is equivalent to a statement about the number of points on the curve.

The usual Riemann zeta function $\zeta(s)$, given by $\sum_{n=1}^{\infty} n^{-s}$ for $\text{Re}(s) > 1$ (and which can be extended to $\mathbb{C} \setminus 1$), satisfies the identity

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

written with pleasing symmetry. A famous conjecture asserts that for $s \in \mathbb{C} \setminus \{-2, -4, \dots\}$ such that $\zeta(s) = 0$, $\mathrm{Re}(s) = \frac{1}{2}$. For an elliptic curve E over $\mathbb{F}_{p^k}$, where we will write $q = p^k$ and $\ell = q + 1 - |G(E, \mathbb{F}_q)|$, we define

$$\zeta_E(s) = \exp\left(\sum_{n=1}^{\infty} \frac{|G(E, \mathbb{F}_{q^n})|}{n} q^{-ns}\right).$$

Theorem 16 ([Wa, pp. 97, 355]). *We have:*

$$\zeta_E(s) = \frac{q^{1-2s} - \ell q^{-s} + 1}{(1 - q^{-s})(1 - q^{1-s})}$$

which has the evident symmetry $\zeta_E(s) = \zeta_E(1 - s)$.

Proof. First consider the polynomial $x^2 - \ell x + q = (x - \rho_1)(x - \rho_2)$. This divides $(x^n - \rho_1^n)(x^n - \rho_2^n) = x^{2n} - (\rho_1^n + \rho_2^n)x^n + q^n$ (since $\rho_1\rho_2 = q$), with quotient we will call $f(x)$.

We have $\varphi_q^2 - \ell\varphi_q + q = 0$ by Theorem 12, so $0 = f(\varphi_q)(\varphi_q^2 - \ell\varphi_q + q) = \varphi_q^{2n} - (\rho_1^n + \rho_2^n)\varphi_q^n + q^n = 0$, so by the uniqueness of ℓ from Theorem 12, $\rho_1^n + \rho_2^n = q^n + 1 - |G(E, \mathbb{F}_{q^n})|$. (It is easy to see inductively that $\rho_1^n + \rho_2^n$ is an integer for all n.) In particular,

$$\zeta_E(s) = \exp\left(\sum_{n=1}^{\infty} \frac{|G(E, \mathbb{F}_{q^n})|}{n} q^{-ns}\right) = \exp\left(\sum_{n=1}^{\infty} \frac{q^n + 1 - \rho_1^n - \rho_2^n}{n} q^{-ns}\right)$$

Using the Taylor expansion $\log(1 - q^{-s}) = \sum -\frac{q^{-ns}}{s}$, we obtain

$$\begin{aligned}\zeta_E(s) &= \exp\left(-\log(1 - q^{1-s}) - \log(1 - q^{-s}) + \log(1 - \rho_1 T) + \log(1 - \rho_2 T)\right) \\ &= \frac{q^{1-2s} - \ell q^{-s} + 1}{(1 - q^{-s})(1 - q^{1-s})}.\end{aligned}$$ □

Moreover, we can verify the Riemann hypothesis in this case.

Theorem 17 ([Wa, p. 357]). *If $\zeta_E(s) = 0$, $\mathrm{Re}(s) = \frac{1}{2}$.*

Proof. Indeed, for the numerator $q^{-2s}(q^{2s} - \ell q^s + q)$ to vanish, we find by the quadratic formula (treating the numerator as a quadratic in q^s) that $q^s = \frac{\ell \pm \sqrt{\ell^2 - 4q}}{2}$. By the Hasse bound, $\ell^2 \leq 4q$, so the two solutions for q^s satisfy $|q^s| = \sqrt{\frac{\ell^2}{4} + \frac{4q - \ell^2}{4}} = \sqrt{q}$. This implies $|q^s| = q^{\mathrm{Re}(s)} = \sqrt{q}$, or $\mathrm{Re}(s) = \frac{1}{2}$. □

2.5 Elliptic Curve Cryptography

In this section we present the factoring algorithm using elliptic curves. After this we study the discrete logarithm problem on elliptic curves and study its security. Assuming the hardness of this problem, we provide elliptic curve variants of several of the cryptosystems mentioned in Section 2.2.3. Finally, we detail the practical implications of using elliptic curves in place of computations over $\mathbb{F}_p^\times$.

2.5.1 Factoring via $G(E, \mathbb{Z}/n\mathbb{Z})$

It is now very easy to describe the elliptic curve factoring algorithm, since it is a minor modification of Pollard's algorithm, and uses in a simple way the development of the preceding two sections. The key difference is that instead of working over $\mathbb{Z}/n\mathbb{Z}$, we work over $G(E, \mathbb{Z}/n\mathbb{Z})$ for a suitably chosen elliptic curve E.

We first randomly choose a curve $y^2 = x^3 + ax + b$ with a point P on it. As Lenstra observes in [Le], this can be done efficiently by choosing $a \in \mathbb{Z}/n\mathbb{Z}$ randomly, choosing a random pair

$P = (u, v) \in (\mathbb{Z}/n\mathbb{Z})^2$, and setting $b = v^2 - u^2 - au \bmod n$, so that P lies on E. Then for some large integer k with many small prime factors, we compute kP by repeated squaring, using the same formula as if $\mathbb{Z}/n\mathbb{Z}$ were a field. (We could, as in Pollard's algorithm, take $k = \operatorname{lcm}(1, \ldots, N)$ for some large integer N.) Note that inverses modulo n are efficiently computable via Euclid's algorithm, so this process is computationally efficient. If at any point during this computation, we fail because an inverse cannot be computed, we have found a nonzero nonunit of $\mathbb{Z}/n\mathbb{Z}$, thus factoring n.

Write $\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}/p_1^{\alpha_1}\mathbb{Z} \times \cdots \times \mathbb{Z}/p_k^{\alpha_k}\mathbb{Z}$, and consider the induced decomposition of the elliptic curve given in Theorem 8. Then the failure to compute an inverse by the usual method occurs whenever $P \in G(E, \mathbb{Z}/p_1^{\alpha_1}\mathbb{Z}) \times \cdots \times G(E, \mathbb{Z}/p_k^{\alpha_k}\mathbb{Z})$ has a coordinate that is the point at infinity in $G(E, \mathbb{Z}/p_i^{\alpha_i}\mathbb{Z})$ for some i. This is essentially the same behavior as in Pollard's algorithm, except that the component elliptic curves have varying orders. The requirement that k have many shared factors with $\varphi(n)$ now changes to having shared factors with the orders of these component curves. By picking many random curves, it is likely that one of the curves will have small factors in $\varphi(n)$.

2.5.2 Elliptic Curve Discrete Logarithm Problem

The elliptic curve discrete logarithm problem is as follows: Given $G(E, K)$, a point $P \in G(E, K)$, and $Q = nP$, compute n. There are a tremendous number of algorithms that aim to solve this problem, many for very specific classes of elliptic curves. A short introduction to this can be found in [Wa] and a thorough treatment can be found in [CFA]. Here we cover general methods developed by Pollard, as well as a solution using the results of Section 2.4 specific to a certain class of curves.

2.5.2.1 Pollard ρ and λ Algorithms

The **Pollard ρ algorithm** is very simple and general; it can be used to compute the discrete logarithm in $\mathbb{F}_p^\times$ as well. We will write the algorithm here using the notation of arithmetic on an elliptic curve, with uppercase letters representing points on the curve and lowercase letters representing integers. Pick a function $f : G(E, K) \to G(E, K)$ of *sets* that, as [Wa] describes, "behaves rather randomly." While f need not be an endomorphism, we will require it to be reasonably explicit, in a sense made precise below. The intuition is that if f were chosen so that its value at every point was another truly random value of the curve, one would expect iterating P, $f(P)$, $f(f(P))$, etc. to repeat a value in roughly $\sqrt{\frac{\pi|G(E,K)|}{2}}$ steps.

Pick random i_0, j_0 and define $S_0 = p_0P + q_0Q$, where $nP = Q$ is the discrete logarithm instance we are trying to solve. Also define $S_i = f(S_{i-1})$, but keep track of values p_i, q_i such that $S_i = p_iP + q_iQ$. This imposes a requirement that f be sufficiently explicit for the image of a point of the form $S_{i-1} = p_{i-1}P + q_{i-1}Q$ to have an efficiently computable representation in the form $p_iP + q_iQ$. In other words, we need to be able to write the difference $f(S_i) - S_i$ in the form $pP + qQ$. Now suppose that for some integers $\alpha \neq \beta$, $S_\alpha = S_\beta$. This will eventually be the case. since $G(E, K)$ is finite. Then $p_\alpha P + q_\alpha Q = p_\beta P + q_\beta Q$, or $P(p_\alpha - p_\beta) = Q(q_\beta - q_\alpha)$. From here, there are only $k = \gcd(q_\beta - q_\alpha, |G(E, K)|)$ different possible choices for the logarithm, so we can test them all. This is because modulo $m = \frac{|G(E,K)|}{k}$, the value of n is simply $(q_\beta - q_\alpha)^{-1}(p_\alpha - p_\beta)$, and there are only k possible choices modulo $|G(E, K)|$ that leave this residue modulo m.

The **Pollard λ algorithm** is a variant on this idea. Instead of having a single starting point S_0, this process is simultaneously carried out on an array of values $S_{0,0}, S_{0,1}, \ldots, S_{0,k}$. If there is a match between two different paths, we can get a relationship similar to that in the ρ algorithm, and again find few possibilities for the logarithm.

Interestingly, as [Wa] explains, the ρ and λ are chosen to match the nature of these algorithms. In the ρ algorithm, one searches for a path that loops back to itself, which might look like the Greek letter ρ. In the λ algorithm, two paths need to converge together, which looks like a λ.

2.5.2.2 Reduction via the Weil Pairing

Starting with a result of Menezes, Okamoto, and Vanstone in [MOV], cryptographers have managed to use Weil and other pairings on elliptic curves to reduce instances of the elliptic curve discrete logarithm problem to instances of the discrete logarithm problem over finite fields, which are usually much easier to solve.

It is possible to efficiently compute the order of a point as a consequence of Theorem 15, which can be used to restrict the range of possible orders of the point to just finitely many, each of which can then be checked. There are faster algorithms, a subject covered in Section 19.4 of [CFA].

Suppose that we need to compute $s = \log_P Q$, where P has order n. We have $G_n(E, \mathbb{F}_p) \subseteq G(E, \mathbb{F}_{p^k})$ for some choice of k. One chooses $R \in G(E, \mathbb{F}_{p^k})$, and computes its order m. Next one computes $R' = \frac{m}{(m,n)} R \in G_n(E, \mathbb{F}_p)$ (since $(m,n) \mid n$), and computes $\ell = \log_{w_n(P,R')} w_n(Q, R')$, a discrete logarithm problem in $\mathbb{F}_{p^k}$. Then $w_n(P, R')^\ell = w_n(\ell P, R') = w_n(Q, R')$, implying that in the subgroup $w_n(G_n(E, \mathbb{F}_p), R') = C_{(m,n)} \subseteq C_n$, we have $\ell P = Q$, which in turn implies that $s \equiv \ell \bmod (m,n)$. For sufficiently many choices of R, the values $\ell \bmod (m,n)$ should allow one to reconstruct $s \bmod n$.

Unfortunately, the discrete logarithm problem over $\mathbb{F}_{p^k}$ is difficult if k is large. We define a curve E over $\mathbb{F}_p$ to be **supersingular** if $|G(E, \mathbb{F}_p)| = p + 1 - a$ where $a \equiv 0 \bmod p$. One can show that k is rather small in the case of such a curve. Following [Wa], we will do this in the special case of $a = 0$.

Theorem 18 ([Wa, p. 146]). *If $|G(E, \mathbb{F}_p)| = p + 1$, and there exists $P \in G(E, \mathbb{F}_p)$ of order n, then we can take $k = 2$ above.*

Proof. Since P has order n, $n \mid p + 1$ by Lagrange's theorem, and thus $1 \equiv -p \bmod n$. Let $Q \in G_n(E, \mathbb{F}_p)$. We have $\varphi_p^2 = -p$ by 12, so $\varphi_p^2(Q) = -pQ = 1 \cdot Q = Q$ since $nQ = 0$. Since φ_p^2 fixes exactly $G(E, \mathbb{F}_{p^2})$, $Q \in G(E, \mathbb{F}_{p^2})$. □

2.5.3 Cryptography

Many of the cryptographic schemes presented in Section 2.2.3 depended upon the hardness of the discrete logarithm problem in $\mathbb{F}_p^\times$. Resting on the assumption that the elliptic curve discrete logarithm problem is hard for a class of instances of $G(E, R)$ that can be efficiently generated, we can develop secure cryptographic protocols. The following are given in [Wa], though the details of their security are discussed more fully in [CFA]. As in the last subsection, we use lowercase letters to indicate integers and uppercase letters to denote points on the elliptic curve. In some cases, such as the Elgamal Public Key Encryption scheme, the message should be encoded as a point on the curve. In others, such as the Digital Signature Algorithm, it is encoded directly as an integer.

2.5.3.1 Diffie-Hellman Key Exchange

It is an open problem to prove the difficulty of determining the point xyP from P, xP, and yP, even under the assumption that the discrete logarithm problem is hard. This problem is similar to that of the Diffie-Hellman problem described in Section 2.2.2, and is called the **elliptic curve Diffie-Hellman problem.**

For the protocol, Alice and Bob agree on a choice of E and p so that the elliptic curve Diffie-Hellman problem is hard for $G(E, \mathbb{F}_p)$, and agree on a point $P \in G(E, \mathbb{F}_p)$. Alice secretly chooses x while Bob secretly chooses y, both integers. Alice sends xP to Bob, who sends yP back. They both compute xyP, and extract a key from it.

2.5.3.2 Elgamal Public Key Encryption

Suppose that Bob wants to receive a message. He publishes an elliptic curve E over $\mathbb{F}_p$, p, a point P, and xP, where x is his secret key. Alice encrypts her message $M \in G(E, \mathbb{F}_p)$ by picking a secret integer y and computing and sending $C_1 = yP$, $C_2 = M + y(xP)$ to Bob, who decrypts by computing $C_2 - xC_1 = M$. It is unclear how an eavesdropper would be able to compute M without solving the discrete logarithm problem, though in the manner of Theorem 1 one can relate the hardness of breaking this cryptosystem to the elliptic curve Diffie-Hellman problem. Again, the modification to produce an elliptic curve algorithm from that given above was simply a matter of writing the variables additively and using the group of points on a curve E.

2.5.3.3 Digital Signature Algorithm

Recall that Alice has a document $m \in \mathbb{Z}$ that she wishes to sign. Alice picks E over $\mathbb{F}_q$, where $|G(E,\mathbb{F}_q)| = ep$ for p a prime and e very small, say $1, 2$, or 4. She also picks a point $P \in G(E,\mathbb{F}_q)$ of order p, just as in the classical variant. Finally she picks a secret integer s and publishes $(E, \mathbb{F}_q, p, e, P, sP)$. Alice picks a random integer $1 \leq \ell < p$, computes $R = \ell P = (x, y)$ and $a = k^{-1}(m + sx) \bmod p$, and produces a signed document (m, R, a). The verifier Bob computes $c_1 = a^{-1}m \bmod p$, $c_2 = a^{-1}x \bmod p$, and accepts if $R = c_1 P + c_2(sP)$. A correctly signed document is always accepted, since $R = kP = a^{-1}(m + sx)P = c_1 P + c_2(sP)$. As before, if the discrete logarithm problem is hard, it seems difficult to use the public information, even with many signed documents, to forge a signature, except perhaps in some special cases. It is unknown, however, whether the hardness of breaking this algorithm follows from the hardness of the discrete logarithm.

2.5.4 Cryptosystems Using the Weil Pairing

In [BF}, Boneh proposed an identity-based encryption scheme using a type of nondegenerate bilinear pairing that can be constructed, for example, from the Weil pairing. In this section, we follow this paper's construction, though for simplicity we provide only the weakest cryptosystem developed (one that is vulnerable to certain attacks). Moreover, we will define the new pairing only for the curve E given by $y^2 = x^3 + 1$.

Let $p \equiv 2 \bmod 3$ be prime. We choose E as above because the pairing needed has a particularly natural description in terms of the automorphism $\sigma : G(E, \mathbb{F}_{p^2}) \to G(E, \mathbb{F}_{p^2})$ given by $(x, y) \mapsto (x, \zeta y)$, where ζ is a primitive third root of unity. One can show, moreover that $|G(E, \mathbb{F}_p)| = p + 1$, so the weakness proved in Theorem 18 applies here.

We fix a prime $q \mid p + 1$ and a point $P \in G(E, \mathbb{F}_p)$ of order q. Then the **modified Weil pairing** is the function $\tilde{w}_q(P_1, P_2) = w_q(P_1, \sigma(P_2))$, though we will be interested in its restriction to $S \times \sigma(S)$, where $S = \langle P \rangle$. The modified pairing, still bilinear, satisfies a different kind of nondegeneracy property: for P' a generator of $G_q(E, \mathbb{F}_p)$, $\tilde{w}_q(P', P')$ is a primitive q^{th} root of unity. We denote by $C_q \subseteq \mathbb{C}$ the group of q^{th} roots of unity.

Setup. A secret integer s is chosen by the **Private Key Generator** (PKG), while $sP = Q$ is made public, together with p, q, and P. Two cryptographic hash functions $H_1 : \{0,1\}^* \to S \setminus \{0\}$ and $H_2 : C_n \to \{0,1\}^m$ for some fixed m are also made public.

Each party comes to collect from the PKG their private key d_B, which the PKG generates from their identifier I_B by the formula $d_B = sH_1(I_B)$.

Encryption. To send a message m to Bob, who has identifier I_B, Alice generates r randomly from C_q, computes the encryption key $e_B = \tilde{w}_q(H_1(I_B), Q)$, and computes the cyphertext $C = (rP, m \oplus H_2(e_B^r))$. She sends C to Bob.

Decryption. Given the cyphertext $C = (c_1, c_2)$, Bob computes

$$\begin{aligned} c_2 \oplus H_2(\tilde{w}_q(d_B, c_1)) &= m \oplus H_2(e_B^r) \oplus H_2(\tilde{w}_q(sH_1(I_B), rP)) \\ &= m \oplus H_2(\tilde{w}_q(H_1(I_B), Q)^r) \oplus H_2(\tilde{w}_q(H_1(I_B), Q)^r) = m \end{aligned}$$

where the second equality is by linearity.

Nondegeneracy is critical in proving the security of this algorithm, which is proved in [BF].

As a final note, a simpler example of the utility of the Weil pairing is in the one-round tripartite matching protocol proposed by Antoine Joux in [Jo]. In this protocol, parties p_1, p_2, p_3 have a public non-degenerate (in the sense defined in this section) bilinear map $b : G \times G \to C_n$ such as the modified Weil pairing above and generator $g \in G$. They pick random secrets s_1, s_2, s_3, and p_i sends $s_i g$ to every other player. Finally p_1 computes $b(s_2 g, s_3 g)^{s_3} = b(g, g)^{s_1 s_2 s_3}$, as do each of the other players. The value of $b(g, g)^{s_1 s_2 s_3}$ is the established common key.

The theory of pairings is one of the most active and interesting in elliptic curve cryptography. Indeed, there is an entire conference each year, called Pairing, dedicated solely to their study. There are additional pairings other than the Weil and modified Weil pairings discussed here, including the Tate and Tate-Lichtenbaum pairings, and many more applications. One can even define pairings on more general varieties than elliptic curves.

2.5.5 Practical Considerations

The value of elliptic curve cryptography, already illustrated by the discussion of pairings in the previous section, is increased by several more pragmatic considerations.

Security. The primary benefit of elliptic curve cryptography is that seemingly greater security may be achieved than for cryptosystems implementing the usual discrete logarithm problem. As explained in [HMV], the most efficient algorithms to solve the discrete logarithm problem over $(\mathbb{Z}/p\mathbb{Z})^\times$ are in subexponential time, far more efficient than the best algorithms to solve the discrete logarithm problem on general elliptic curves (Pollard's algorithm, discussed in Section 2.5.2.1, is essentially the best known method).

Efficiency. The Elgamal cryptosystem based on the usual discrete logarithm problem was discussed in the section on cryptography. This algorithm requires computations that in practice are far more time consuming than the equivalent computations on elliptic curves, where doubling and addition tend to be more efficient. Moreover, the security benefit mentioned earlier creates a new source of efficiency. The integers needed to do elliptic cryptography with comparable security to classical cryptosystems are usually a tenth of the size, creating a significant speedup in arithmetic on the curve.

Versatility. Elliptic curves, as illustrated by the factoring algorithm, are numerous for a given choice of ring $\mathbb{Z}/n\mathbb{Z}$. This creates an extra dimension of versatility and functionality missing in classical algorithms, and explains the huge improvement over Pollard's factoring methods.

Structure. The Weil pairing and other structures defined on elliptic curve provide a rich range of tools. As Sections 2.5.2.2 and 2.5.4 reveal, the interaction of these structures on the curve can be helpful for designing cryptosystems or performing computational tasks.

It is not wise to immediately replace all one's cryptosystems by those employing elliptic curve methods, however. Elliptic curves are very complex objects, and in most algorithms above, the curve E is published. The discrete log problem has been shown to be much easier on many subclasses of elliptic curves by results such as Theorem 18, so if one selects a curve with particular properties, many of which might not be readily apparent, an adversary could crack the cryptosystem using "special" techniques specific to that curve. That said, much work has indeed been done on the selection process to produce a curve with optimal security, and [CFA] and [HMV] provide substantial coverage of this.

The study of more sophisticated number theory such as modular forms and complex multiplication can have cryptographic implications, as discussed in [CFA]. A generalization of elliptic curve cryptography is hyperelliptic curve cryptography, which performs arithmetic on the Jacobian of hyperelliptic curves. Elliptic curve cryptography is a subject where deep number theory has direct impact on practical matters, and the results described in Sections 2.4 and 2.5 only scratch the surface.

2.6 Conclusions

For the reader interested in investigating this material further, the following books provide additional information on the topics discussed in this paper:

Cryptography. The approach taken by this paper most closely parallels the philosophy of [MvOV], which focuses on providing descriptions of real cryptosystems, rather than establishing the foundations of the subject. An amazing introduction to foundational cryptography can be found in [Go1] and [Go2].

Elliptic Curves. Three introductory texts on elliptic curves are [M2], [ST], and [Wa], which together cover all the material here and much more. For a more advanced introduction, look to [Si1] and [Si2] or [Hu]. For those interested in the complex-analytic aspects of the subject, [La] and [Fo] are excellent texts. From an algebro-geometric viewpoint, one might look into [Ha] or [M1].

Elliptic Curve Cryptography. For all-purpose text that is encyclopedic in both its coverage of the theory and practice of elliptic curve cryptography, look to [CFA]. The text [HMV] focuses on the practical aspects of the subject and is very readable.

The group of points on an elliptic curve can be used as a replacement for multiplicative groups in many situations. As we saw in the factoring algorithm, the key improvement was that factoring n via Pollard's algorithm could be reduced to "guessing" a number that is either a multiple of $\varphi(n)$, or has many factors in common with it, while over an elliptic curve, $\varphi(n)$ is replaced with the group order, which can take on many values for fixed n as E varies. This added flexibility gives a new dimension to the development of cryptographic systems, since, as seen in the examples above, one can vary the curve E in setting parameters while keeping the field fixed. Although we did not delve into the more complex algorithms that make more important use of this phenomenon, the factoring algorithm shows that elliptic curves can provide important computational savings. We also saw that structures such as the Weil pairing can give rise to cryptosystems for which no simple implementation using classical methods are known. For their security, efficiency, and versatility, elliptic curve methods are used for real cryptographic applications every day, revealing their amazing pervasiveness in both theory and applications within mathematics and computer science.

References

[BF] Dan Boneh and Matt Franklin: Identity-Based Encryption from the Weil Pairing, *Advances in Cryptology-Crypto 2001: 21st Annual International Cryptology Conference, Santa Barbara, California, USA, August 19-23, 2001, Proceedings* (2001).

[Bo] Dan Boneh: Twenty years of attacks on the RSA cryptosystem, *Notices of the Amer. Math. Soc.* **46** #2 (1999), 203–213.

[CFA] Henri Cohen, Gerhard Frey, and Roberto Avanzi: *Handbook of Elliptic and Hyperelliptic Curve Cryptography.* Boca Raton: CRC Press 2006.

[CLRS] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein: *Introduction to Algorithms.* Cambridge, MA: MIT Press, McGraw-Hill 2001.

[Fo] Otto Forster: *Lectures on Riemann Surfaces.* New York: Springer-Verlag 1981.

[Ga] Steven D. Galbraith: The Weil Pairing on Elliptic Curves over $\mathbb{C}$, preprint (2005).

[GB] Shafi Goldwasser and Mihir Bellare: Lecture notes on cryptography, *Summer Course, "Cryptography and Computer Security," at MIT* 1996.

[Go1] Oded Goldreich: *Foundations of Cryptography: Volume 1, Basic Tools.* New York: Cambridge Univ. Press 2001.

[Go2] Oded Goldreich: *Foundations of Cryptography: Volume 2, Basic Applications.* New York: Cambridge Univ. Press 2004.

[Ha] Robin Hartshorne: *Algebraic Geometry.* New York: Springer-Verlag 1977.

[HMV] Darrel R. Hankerson, Alfred J. Menezes, and Scott A. Vanstone: *Guide to Elliptic Curve Cryptography.* New York: Springer-Verlag 2004.

[Hu] Dale Husemoller *Elliptic curves.* New York: Springer-Verlag 2004.

[Jo] Antoine Joux: A One Round Protocol for Tripartite Diffie–Hellman, *J. of Cryptology* **17** #4 (2004), 263–276.

[Ko] Neal I. Koblitz: *A Course in Number Theory and Cryptography.* New York: Springer-Verlag 1994.

[La] Serge Lang: *Elliptic Functions.* New York: Springer-Verlag 1987.

[Le] Hendrik W. Lenstra: Elliptic curves and number-theoretic algorithms, (1986).

[MOV] Alfred J. Menezes, Tatsuaki Okamoto, and Scott A. Vanstone: Reducing elliptic curve logarithms to logarithms in a finite field, *Information Theory, IEEE Transactions on*, **39** #5 (1993),1639–1646.

[MvOV] Alfred J. Menezes, Paul C. van Oorschot, and Scott A. Vanstone: *Handbook of Applied Cryptography*. Boca Raton: CRC Press 1997.

[M1] James S. Milne: Abelian varieties, course notes.

[M2] James S. Milne: Elliptic curves, course notes.

[Shaf] Igor R. Shafarevich: *Basic Algebraic Geometry, Volume 1*. New York: Springer-Verlag 1994.

[Sham] Adi Shamir: Identity-based cryptosystems and signature schemes, *Proceedings of CRYPTO 84* (1985), 47–53.

[Si1] Joseph H. Silverman: *The Arithmetic of Elliptic Curves*. New York: Springer-Verlag 1986.

[Si2] Joseph H. Silverman: *Advanced Topics in the Arithmetic of Elliptic Curves*. New York: Springer-Verlag 1994.

[ST] Joseph H. Silverman and John Tate: *Rational Points on Elliptic Curves*. New York: Springer-Verlag 1992.

[Si] Michael Sipser: *Introduction to the Theory of Computation*. Course Technology 1996.

[Wa] Lawrence C. Washington: *Elliptic Curves: Number Theory and Cryptography*. Boca Raton: Chapman & Hall/CRC Press 2003.

STUDENT ARTICLE

3
Bridging the Group Definition Gap

Matthew G. Dawson[†]
Union University '08
Jackson, TN.
s285618@uu.edu
abc285618@gmail.com

Abstract

In the early 1830s, a young French mathematician named Évariste Galois laid the foundations of group theory, although he never precisely defined groups. Galois studied groups in the context of sets of arrangements and his ideas were reformulated into a more abstract setting in the twentieth century. This paper provides precise definitions for constructs closely related to Galois's original notion of group theory and explores important group properties in that context, demonstrating that the modern concepts of the group, subgroup, normal subgroup, and solvable group can be expressed in terms of arrangement sets.

In the early nineteenth century, the theory of polynomials in a single variable was significantly advanced. Paolo Ruffini (1765–1822) discovered that the general quintic polynomial is not solvable by radicals and produced a nearly complete proof of this result. Niels Abel (1802–1829) was able to solve one of the greatest open questions of his day when he provided a correct and complete proof of Ruffini's discovery. A precocious French mathematician by the name of Evariste Galois (1811–1832) then made a surprisingly complete advancement, discovering a criterion that determines when a polynomial is solvable by radicals. Along the way, he stumbled upon the branch of mathematics now known as group theory. Galois associated groups with polynomial equations, showing that a polynomial equation is solvable by radicals when the associated group has a certain property now known as **solvability**.

Interestingly, although Galois was the first to study groups in the abstract setting, his concept of group bears no superficial resemblance to the more familiar definition that is found in modern textbooks. Today, a **group** is defined as a set, together with an associative binary operation, having both the **identity** and **inverse** properties.[1] Galois, however, thought of groups in the context of **arrangements**. A rigorous study of Galois's arrangement sets and their relation to modern group theory is not well known and is difficult to find, although such an exposition may be found in [Ti]. This paper reconciles the two definitions of group and studies the relationships between the two perspectives. Ultimately, we shall determine precisely how the modern definition of solvable group translates into Galois's terminology.

To begin our journey, we must provide precise definitions for the terminology that Galois used. First, we define the concept of arrangement key to Galois's formulation of group theory.

Definition 1. Given a nonempty finite set S of n elements, an **arrangement** of S is an n-tuple $(a_1, a_2, \ldots, a_n) \in S^n$ such that for every element $s \in S$ there exists exactly one i such that $a_i = s, 1 \leq i \leq n$.

In addition, the set of all arrangements of a set S is denoted by $\mathrm{Arr}(S)$ and the set of all permutations on S is denoted by $\mathrm{Sym}(S)$.

[†] Matthew Dawson, Union University '08, is a mathematics major who lives in Jackson, TN. Thanks to home education from his parents, he will be graduating four years early. In addition to mathematics, interests include physics, computer science, history, and music.

[1] That is, there is an identity element of the group and every element of the group has an inverse.

To illustrate this definition, let us consider a simple example. Suppose $S = \{a, b, c\}$. We list the elements of $\text{Arr}(S)$, denoting abc rather than (a, b, c) for brevity:

$$\text{Arr}\left(\{a, b, c\}\right) = \{abc, acb, bac, bca, cab, cba\}.$$

A **permutation** on S is simply a one-to-one correspondence mapping S into itself. Using the cyclic notation for permutations, we have that $\text{Sym}(S) = \{(ab), (bc), (ac), (abc), (acb), \text{id}\}$.

The power and versatility of arrangements is demonstrated by our first observation, which allows us to associate permutations on $\text{Arr}(S)$ with permutations on S.

Proposition 2. *Let S be a finite set with n elements.*

1. *Let $f \in \text{Sym}(S)$, and consider the mapping P_f on $\text{Arr}(S)$ such that for each arrangement $\alpha = (a_1, a_2, a_3, \ldots, a_n) \in \text{Arr}(S)$,*

$$P_f(\alpha) = (f(a_1), f(a_2), f(a_3), \ldots, f(a_n)).$$

   Then P_f is a permutation on $\text{Arr}(S)$.

2. *For all $\alpha, \beta \in \text{Arr}(S)$, there exists a unique permutation $f \in \text{Sym}(S)$ such that $P_f(\alpha) = \beta$.*

3. *For all $f, g \in \text{Sym}(S)$, $P_f \circ P_g = P_{f \circ g}$.*

The third part of Proposition 2 establishes that the map $f \mapsto P_f$ is a homomorphism from $\text{Sym}(S)$ to $\{P_f \mid f \in \text{Sym}(S)\} \subset \text{Sym}(\text{Arr}(S))$; the second part tells us that it is an isomorphism; that is, $P_f = P_g$ if and only if $f = g$.

The fact that such an isomorphism exists should not be surprising, when Cayley's Theorem is considered: clearly, if a set S has n elements, then $|\,\text{Sym}(S)| = n!$ and $|\,\text{Arr}(S)| = n!$. Cayley's theorem tells us that because $\text{Sym}(S)$ has $n!$ elements, it will be isomorphic to a subgroup of $S_{n!}$. Since $\text{Sym}(\text{Arr}(S))$ is isomorphic to $S_{n!}$, we see that there must be an isomorphism between $\text{Sym}(S)$ and some subgroup of $\text{Sym}(\text{Arr}(S))$.

Proposition 2 assures us that the permutations in $\text{Sym}(S)$ can be applied to arrangements in $\text{Arr}(S)$ in a well-behaved fashion. Indeed, Proposition 2 sets up a group action of $\text{Sym}(S)$ on $\text{Arr}(S)$. It does so because the map $f \mapsto P_f$ is a homomorphism from $\text{Sym}(S)$ to $\text{Sym}(\text{Arr}(S))$. Furthermore, the second part of Proposition 2 implies that the homomorphism is one-to-one, so that the group action is **faithful**.

Corollary 3. *Let S be a finite set. Then the mapping $P : \text{Sym}(S) \to \text{Sym}(\text{Arr}(S))$ given by $P(f) = P_f$ is a faithful action of* $\text{Sym}(S)$ *on* $\text{Arr}(S)$.

Henceforth, we will drop the notation P_f and instead use the same notation to denote a permutation on S and the corresponding permutation on $\text{Arr}(S)$. In addition, we will denote function composition by juxtaposition.

Now that a group action has been set up, the concept of orbit may be discussed. The reader may recall that, given a group G, a set M, and an action of G on M, the orbit of $m \in M$ is defined to be $G(m) = \{g(m) \mid g \in G\}$.

Now, if a set S and an arrangement $\alpha \in \text{Arr}(S)$ are considered, then the orbit of α is $\text{Arr}(S)$ (that is, $(\text{Sym}(S))(\alpha) = \text{Arr}(S)$). We know this by part two of Proposition 2, which tells us that given any arrangement in $\text{Arr}(S)$, we can find a permutation in $\text{Sym}(S)$ that will map α to that arrangement.

A more general concept similar to that of orbit will be quite useful in this paper; we need to consider the application of subsets of $\text{Sym}(S)$ on a single arrangement.

Definition 4. Let S be a nonempty finite set and let $H \subseteq \text{Sym}(S)$. Then for all arrangements $\alpha \in \text{Arr}(S)$, we define $H(\alpha) = \{f(\alpha) \mid f \in H\}$.

To illustrate this definition, let $H = \{\mathrm{id}, (abc), (acb), (ac)\}$. In this case, we have

$$H(abc) = \{abc, bca, cab, cba\}.$$

Now suppose we are given a finite set S, an arrangement α of S, and a set M of arrangements of S. We define:

Definition 5. Let S be a nonempty finite set, let $C \subseteq \mathrm{Arr}(S)$, and let $\alpha \in C$. Then the **permutation set** of α in C, denoted $\bowtie_\alpha(C)$, is the set

$$\bowtie_\alpha(C) = \{f \in \mathrm{Sym}(S) \mid f(\alpha) \in C\}.$$

Let us go back to our previous example, where $S = \{a, b, c\}$, $C = \{abc, bca, cab, cba\}$, and $\alpha = abc$. Then $\bowtie_\alpha(C)$ will be the set of all permutations that map abc to some arrangement in C. The reader can check that $\bowtie_\alpha(C) = \{\mathrm{id}, (abc), (acb), (ac)\}$.

As suggested above, the $\bowtie_\alpha(C)$ construction is the inverse of the $H(\alpha)$ construction. We state this formally in the following lemma.

Lemma 6. *Let S be a nonempty finite set, let $H \subseteq \mathrm{Sym}(S)$ and $M \subseteq \mathrm{Arr}(S)$, and let $\alpha \in M$. Then $H(\alpha) = M$ if and only if $H = \bowtie_\alpha(M)$.*

Proof. First suppose that $H(\alpha) = M$. We wish to show that $H = \bowtie_\alpha(M)$. Let $f \in H$. Then $f(\alpha) \in H(\alpha) = M$ and hence $f \in \bowtie_\alpha(M)$. Thus, $H \subseteq \bowtie_\alpha(M)$.

Next let $h \in \bowtie_\alpha(M)$. Thus $h(\alpha) \in M = H(\alpha)$, so that $h(\alpha) = f(\alpha)$ for some $f \in H$. Therefore, by Proposition 2, we know that $h = f$, so that $h \in H$. Hence, $\bowtie_\alpha(M) \subseteq H$. Thus, $H = \bowtie_\alpha(M)$.

To prove the other half of the biconditional, suppose that $H = \bowtie_\alpha(M)$. We must show that $H(\alpha) = M$. Let $\beta \in H(\alpha)$, so that $\beta = f(\alpha)$ for some $f \in H = \bowtie_\alpha(M)$. Now $f \in \bowtie_\alpha(M)$ implies that $\beta = f(\alpha) \in M$. Thus $H(\alpha) \subseteq M$.

Finally let $\beta \in M$. Then by Proposition 2, $\beta = g(\alpha)$ for some $g \in \mathrm{Sym}(S)$. Clearly $g \in \bowtie_\alpha(M) = H$. Thus, $g \in H$ implies that $\beta = g(\alpha) \in H(\alpha)$. □

So far, we have associated a permutation set with each pair (α, M) where α is an arrangement and M is an arrangement set. We can also define a permutation set directly associated to a given arrangement set.

Definition 7. Let S be a nonempty finite set, and let $C \subseteq \mathrm{Arr}(S)$. Then the **total permutation set** associated with C (or total permutation set of C), $\bowtie(C)$, is the set

$$\bowtie(C) = \{f \in \mathrm{Sym}(S) \mid \exists \alpha \in C \text{ such that } f(\alpha) \in C\}.$$

By checking the relevant definitions, we see that

$$\bowtie(C) = \bigcup_{\alpha \in C} \bowtie_\alpha(C).$$

Hence, we have that $\bowtie_\alpha(C) \subseteq \bowtie(C)$ for all $\alpha \in C$.

As before, an example will help to illustrate the definition. We consider again $C = \{abc, bca, cab, cba\}$. Then,

$$\begin{aligned} \bowtie(C) &= \bowtie_{\mathrm{abc}}(C) \cup \bowtie_{\mathrm{bca}}(C) \cup \bowtie_{\mathrm{cab}}(C) \cup \bowtie_{\mathrm{cba}}(C) \\ &= \{\mathrm{id}, (abc), (acb), (ac), (ab), (bc)\} = \mathrm{Sym}(S). \end{aligned}$$

Now that we can associate permutation sets with arrangement sets, we are ready to study the implications of those associated permutation sets having special properties, the most important of which is described in our next definition.

Definition 8. A set C of arrangements of a nonempty finite set S is a **Galois Set of Arrangements** (or **(GSA)**) if for all $f \in \bowtie(C)$, $\alpha \in C$ implies $f(\alpha) \in C$.

The above definition lays out the single most important concept in this paper, which is a close approximation to Galois's original concept of the group. Recall that the more familiar definition states that a group is a set of objects with an associative binary operation such that the set has an identity element and contains inverses for every element in the set. Galois sets of arrangements bear no immediate resemblance to this algebraic structure. However, these two concepts are closely related. We shall soon see that the connection between GSAs and algebraic groups arises through the permutation sets associated with GSAs.

Before moving any further, let us determine whether $C = \{abc, bca, cab, cba\}$ is a GSA. First recall that $\bowtie(C) = \{\mathrm{id}, (abc), (acb), (ac), (ab), (bc)\} = \mathrm{Sym}(S)$. In order for C to be a GSA, it must be the case that each permutation $g \in \bowtie(C)$ maps every arrangement in C to another arrangement in C. Consider $f = (ab)$, which is an element of $\bowtie(C)$. Now $f(abc) = bac$, and $bac \notin C$. Therefore, because $f = (ab) \in \bowtie(C)$ yet $abc \in C$ and $f(abc) \notin C$, we see that C cannot be a GSA.

Let us consider another example: suppose that $M = \{abc, acb\}$. We shall first list out all of the elements of $\bowtie(M)$.

$$abc \xrightarrow{\mathrm{id}} abc \qquad abc \xrightarrow{(bc)} acb \qquad acb \xrightarrow{(bc)} abc \qquad acb \xrightarrow{\mathrm{id}} acb$$

Hence $\bowtie(M) = \{\mathrm{id}, (bc)\}$. Because all the permutations in $\bowtie(M)$ map both abc and acb to either abc or acb (this fact can be checked by examining the above diagram), we have that M is a GSA.

The reader may have noticed that $\bowtie_{\mathrm{abc}}(M) = \bowtie_{\mathrm{acb}}(M) = \bowtie(M)$. The reader may also have noticed that $\bowtie_{\mathrm{abc}}(M)$ forms a group of permutations in the modern sense. These observations lead us to an interesting, general result.

Lemma 9. *If C is a set of arrangements of a finite set and $\alpha \in C$ is such that $\bowtie_\alpha(C)$ forms a group under composition, then $\bowtie_\alpha(C) = \bowtie(C)$.*

Proof. Suppose that $\alpha \in C$ such that $\bowtie_\alpha(C)$ forms a group under composition. First we show that $\bowtie(C) \subseteq \bowtie_\alpha(C)$. Let $f \in \bowtie(C)$. Then by definition of the associated permutation set, there exist $\beta, \gamma \in C$ such that $f(\beta) = \gamma$. By Proposition 2, there exists exactly one permutation $g \in \mathrm{Sym}(S)$ such that $g(\alpha) = \beta$, and there exists exactly one permutation $h \in \mathrm{Sym}(S)$ such that $h(\alpha) = \gamma$. Note that by the definition of the permutation set of α in C, $g \in \bowtie_\alpha(C)$ and $h \in \bowtie_\alpha(C)$. Consider the permutation hg^{-1}:

$$(hg^{-1})(\beta) = h(g^{-1}(\beta)) = h(g^{-1}(g(\alpha))) = h(\alpha) = \gamma.$$

By Proposition 2, there exists exactly one permutation f such that $f(\beta) = \gamma$. Thus we have $f = hg^{-1}$. But because $\bowtie_\alpha(C)$ forms a group under composition, we know that $hg^{-1} \in \bowtie_\alpha(C)$. Hence $f \in \bowtie_\alpha(C)$. Thus, $\bowtie(C) \subseteq \bowtie_\alpha(C)$.

By the definition of $\bowtie(C)$, it is clear that $\bowtie_\alpha(C) \subseteq \bowtie(C)$. Therefore, we have that $\bowtie_\alpha(C) = \bowtie(C)$. □

With this last result, we have developed all of the necessary tools to establish the connection between groups and GSAs.

Theorem 10. *Let S be a nonempty finite set, let $C \subseteq \mathrm{Arr}(S)$, and let $\alpha \in C$. Then C is a Galois Set of Arrangements if and only if $\bowtie_\alpha(C)$ forms a group under composition.*

Proof. Suppose that C is a Galois Set of Arrangements. Then, for all $f \in \bowtie(C)$, $f(\beta) \in C$ for all $\beta \in C$. We wish to show that $\bowtie_\alpha(C)$ forms a group with respect to function composition.

Let $f, g \in \bowtie_\alpha(C)$. Thus $g(\alpha) \in C$. Also, C is a GSA, so that $f(\beta) \in C$ for all $\beta \in C$. It follows that $(fg)(\alpha) = f(g(\alpha)) \in C$. Therefore, by the definition of the permutation set of α in C, $fg \in \bowtie_\alpha(C)$. Hence $\bowtie_\alpha(C)$ is closed under composition.

Now consider the identity permutation $\mathrm{id} : S \to S$. Then $\mathrm{id}(\alpha) = \alpha$, so that $\mathrm{id} \in \bowtie_\alpha(C)$. Recall that, since id is the identity permutation, $\mathrm{id} \circ f = f \circ \mathrm{id} = f$ for all permutations $f \in \mathrm{Sym}(S)$. Therefore, the set $\bowtie_\alpha(C)$ contains an identity element.

Next let $f \in \bowtie_\alpha(C)$. Then by the definition of the permutation set of α in C, $f(\alpha) = \gamma$ for some $\gamma \in C$. Now $f^{-1}(\gamma) = \alpha$ (recall that f is a permutation, so that f^{-1} exists), so that

$f^{-1} \in \bowtie(C)$. Thus, since C is a GSA, $f^{-1}(\beta) \in C$ for all $\beta \in C$. Hence, $f^{-1}(\alpha) \in C$ so that $f^{-1} \in \bowtie_\alpha(C)$. Thus, $\bowtie_\alpha(C)$ contains an inverse for each element, whence $\bowtie_\alpha(C)$ forms a group with respect to function composition.

By Lemma 9 we know that $\bowtie_\alpha(C) = \bowtie(C)$. We wish to show that C is a GSA. Let $f \in \bowtie(C)$. In order to show that C is a GSA, we must to show that $f(\beta) \in C$ for all $\beta \in C$. Now $f \in \bowtie_\alpha(C)$, since $\bowtie_\alpha(C) = \bowtie(C)$. Next let $\beta \in C$. By Proposition 2, there exists exactly one permutation $h : S \to S$ such that $h(\alpha) = \beta$. Clearly, $h \in \bowtie_\alpha(C)$. But $\bowtie_\alpha(C)$ forms a group under composition, so that $fh \in \bowtie_\alpha(C)$. In other words, $(fh)(\alpha) = f(h(\alpha)) = f(\beta) \in C$. Therefore, C is a GSA. □

Let us look at Theorem 10 in light of the examples we have used so far. For the set $C = \{abc, bca, cab, cba\}$, we recall that $\bowtie_{\mathrm{abc}}(C) = \{\mathrm{id}, (abc), (acb), (ac)\}$. Now, $\bowtie_{\mathrm{abc}}(C)$ is not a group. Therefore, Theorem 10 tells us that C is not a GSA, confirming our earlier observation. Also, for $M = \{abc, acb\}$, we saw that $\bowtie_{\mathrm{abc}}(M)$ is a group. Theorem 10 then tells us that M is a GSA, as we determined above.

It should be noted that Theorem 10 finishes the task of reconciling the two group definitions. A set C of arrangements is a Galois set of arrangements if and only if at least one of the permutations sets of an arrangement in C is a group. Theorem 10 also implies that if C forms a GSA, then $\bowtie_\alpha(C)$ forms a group for each $\alpha \in C$.

Now, we suppose that $\bowtie_\alpha(C)$ forms a group with respect to function composition. Lemma 9 then guarantees that $\bowtie_\alpha(C) = \bowtie(C)$, so that $\bowtie(C)$ is a group. Thus, if C is a GSA, then the total associated permutation set of C is a group. The converse, however, is not true—the total permutation set of C may be a group even if C is not a GSA. For instance, we showed that $C = \{abc, bca, cab, cba\}$ is not a GSA and also determined that $\bowtie(C) = \mathrm{Sym}(S)$. From group theory, we know that $\mathrm{Sym}(S)$ is a group that is isomorphic to S_3.

Our next priority is to determine how **solvability** translates to the language of permutation sets. Before doing that, we state a few more results and give one more definition.

Lemma 11. *Let S be a nonempty finite set, let $H \subseteq \mathrm{Sym}(S)$, and let M be a GSA of S. Then for all $\alpha \in M$, $H(\alpha) = M$ if and only if $H = \bowtie(M)$.*

Lemma 12. *Let T and V be sets of permutations of a finite set S, and let α be an arrangement of S. Then $T(\alpha) = V(\alpha)$ if and only if $T = V$.*

We have already studied how to apply a permutation set to an arrangement. Next, we shall define the application of a single permutation to an arrangement set.

Definition 13. Let S be a nonempty finite set, and let $M \subseteq \mathrm{Arr}(S)$. Then for all permutations $g \in \mathrm{Sym}(S)$, we define $g(M) = \{g(\gamma) \mid \gamma \in M\}$.

To illustrate the above definition, let $M = \{abc, acb\}$ and $g = (ab)$ and consider the following diagram:

$$abc \overset{(ab)}{\longrightarrow} bac$$

$$acb \overset{(ab)}{\longrightarrow} bca$$

Then, we see that $g(M) = \{bac, bca\}$.

In what follows, we respectively denote gH and Hg for the **left and right cosets** of $H \subset G$ in G associated to $g \in G$.

Theorem 14. *Let N be a GSA of a finite set S, and let $H = \bowtie(N)$. Then for all permutations $g \in \mathrm{Sym}(S)$,*

1. *The set $g(N)$ is a GSA of S and $\bowtie(g(N)) = gHg^{-1}$*

2. *For all $\alpha \in N$, $g(N) = g(H(\alpha)) = (gH)(\alpha)$.*

Proof. First we prove the first statement. Let $\alpha \in N$ and let $g \in \mathrm{Sym}(S)$. Note that because N is a GSA of S, Theorem 11 assures us that $N = H(\alpha)$. Let $\beta = g(\alpha)$. We first show that $\bowtie_\beta(g(N)) = gHg^{-1}$.

Let $k \in \bowtie_\beta(g(N))$. Thus, $k(\beta) = \gamma$ for some $\gamma \in g(N)$. Now, $\gamma = g(\delta)$ for some $\delta \in N$. But $N = H(\alpha)$, so that $\delta = h(\alpha)$ for some $h \in H$. Thus $k(\beta) = g(h(\alpha))$. But $\alpha = g^{-1}(\beta)$, whence $k(\beta) = g(h(g^{-1}(\beta)))$. Therefore, by Proposition 2, we have $k = ghg^{-1}$. Hence, $\bowtie_\beta(g(N)) \subseteq gHg^{-1}$.

Next let $h \in H$ and consider $ghg^{-1} \in gHg^{-1}$. Then $(ghg^{-1})(\beta) = (ghg^{-1})(g(\alpha)) = g(h(\alpha)) \in g(N)$. Hence, $ghg^{-1} \in \bowtie_\beta(g(N))$. Thus, $gHg^{-1} \subseteq \bowtie_\beta(g(N))$.

Therefore, $\bowtie_\beta(g(N)) = gHg^{-1}$. As H is a group under composition, so is gHg^{-1}. Thus, by Theorem 10, $g(N)$ is a GSA. Now, we suppose that $\bowtie_\alpha(C)$ forms a group with respect to function composition. By Lemma 9 we know that $\bowtie(g(N)) = gHg^{-1}$.

Next, we prove the second component of the theorem. Let $g \in \mathrm{Sym}(S)$ and $\alpha \in N$. By Theorem 11, $N = H(\alpha)$, so that $g(N) = g(H(\alpha))$. It remains to show that $g(N) = (gH)(\alpha)$. Thus, suppose $h \in H$ and consider $gh \in gH$. Note that $(gh)(\alpha) = g(h(\alpha))$ and that $h(\alpha) \in N$. Thus, $g(h(\alpha)) \in g(N)$, whence $(gH)(\alpha) \subseteq g(N)$. Next, let $\beta \in g(N)$. Then, $\beta = g(\gamma)$ for some $\gamma \in N$ and so $\gamma = h(\alpha)$ for some $h \in H$. Thus, $\beta = g(h(\alpha)) = (gh)(\alpha)$, whence $g(N) \subseteq (gH)(\alpha)$, whereby $g(N) = (gH)(\alpha)$. □

A few simple observations follow directly from Theorem 14. First, applying a permutation to a GSA yields another GSA. Next, the associated permutation set of the resulting GSA is a conjugate group of the associated permutation set of the original GSA. Finally, in the second part of Theorem 14, we see that the GSA obtained by applying a single permutation to a GSA can be obtained by applying a left coset of the original GSA's associated permutation set to an arrangement in the original GSA.

Thus, the application of a permutation to a GSA results in a GSA that captures the notions of conjugate groups and left cosets. Rigatelli [Ri, p. 124] noticed that Galois referred to cosets as "groups," suggesting that the use of this word might confuse people attempting to understand his work. Perhaps Theorem 14 sheds some light on this issue: the application to an arrangement of the left coset of a permutation group results in a GSA, which is what Galois would call a "group."

We also state a result similar to Theorem 14, which relates right cosets to the application of a permutation group to an arrangement.

Theorem 15. *Let N be a GSA of a finite set S, let $H = \bowtie(N)$, and let $\alpha \in N$. Then for all $g \in \mathrm{Sym}(S), H(g(\alpha)) = (Hg)(\alpha)$.*

Before we can study solvability, we must first study **normal subgroups**. To that end, we shall refine our focus and work more specifically towards establishing a criterion which determines when a given GSA has an associated permutation set that is a normal subgroup of the permutation set of another GSA.

A question immediately arises: if one GSA is a subset of another GSA, is the permutation set of the former GSA a subgroup of the permutation set of the latter GSA? That is, suppose that M and N are GSAs of a finite set S, $N \subseteq M$, and let $G = \bowtie(M)$ and $H = \bowtie(N)$. One would expect that, since $N \subseteq M$, it must be true that $H \leq G$ as groups. To establish the truth of this statement, suppose that $f \in H$. Then, there exists $\alpha \in N$ such that $f(\alpha) \in N$. Since $N \subseteq M$, we have that $f(\alpha) \in M$. Thus, by definition of total associated permutation set, $f \in G$. Thus, $H \subseteq G$. Because M and N are GSAs, we know by Lemma 9 and Theorem 10 that G and H form groups. Therefore, $H \leq G$.

In order to establish a criterion that establishes when GSAs have associated permutation sets where one is a normal subgroup of the other, we need three lemmas and an intermediate theorem. The proofs of the lemmas are straightforward.

Lemma 16. *Let M and N be GSAs of a finite set S, $N \subseteq M$, and let $G = \bowtie(M)$ and $H = \bowtie(N)$. Then for all $g \in G$ and $\alpha \in N$, $(Hg)(\alpha) \subseteq M$ and $(gH)(\alpha) \subseteq M$.*

Lemma 17. *Let M and N be GSAs of a finite set S such that $N \subseteq M$, let $G = \bowtie(M)$ and $H = \bowtie(N)$, and let $\alpha \in N$. Then, the following equalities hold:*

1. $\{H(\beta) \mid \beta \in M\} = \{(Hg)(\alpha) \mid g \in G\}$
2. $\{g(N) \mid g \in G\} = \{(gH)(\alpha) \mid g \in G\}$.

Lemma 18. *Let S be a finite set, let $A \subset \mathrm{Arr}(S)$ and $B \subset \mathrm{Arr}(S)$, and let $\alpha \in \mathrm{Arr}(S)$. Then $(A \cap B)(\alpha) = A(\alpha) \cap B(\alpha)$ and $(A \cup B)(\alpha) = A(\alpha) \cup B(\alpha)$.*

Theorem 19. *Let M and N be GSAs of a finite set S, $N \subseteq M$, and let $G = \bowtie(M)$ and $H = \bowtie(N)$. Then, the following sets are partitions of M:*

$$P_R(M,N) = \{H(\beta) \mid \beta \in M\},$$
$$P_L(M,N) = \{g(N) \mid g \in G\}.$$

These sets are known, respectively, as the ***right and left partitions of*** *M* ***by*** *N.*

Proof. We must show that the sets $P_R(M,N)$ and $P_L(M,N)$ are partitions of M. By Lemma 17, we know that for all $\alpha \in N$, $P_R(M,N) = \{(Hg)(\alpha) \mid g \in G\}$ and $P_L(M,N) = \{(gH)(\alpha) \mid g \in G\}$.

We know from group theory that $\{gH \mid g \in G\}$ forms a partition of G and that $\{Hg \mid g \in G\}$ forms a partition of G. Hence, for all $f, g \in G$, $fH \cap gH = \varnothing$ and $Hf \cap Hg = \varnothing$. Also, $G = \cup_{g \in G} gH$ and $G = \cup_{g \in G} Hg$.

Let $f, g \in G$. Then, we have by Lemma 18 that $(fH)(\alpha) \cap (gH)(\alpha) = (fH \cap gH)(\alpha) = (\varnothing)(\alpha) = \varnothing$. Similarly, $(Hf)(\alpha) \cap (Hg)(\alpha) = (Hf \cap Hg)(\alpha) = (\varnothing)(\alpha) = \varnothing$. Hence, $P_L(M,N)$ and $P_R(M,N)$ are pairwise disjoint.

Finally, Lemma 18 implies that $\cup_{g \in G}((gH)(\alpha)) = (\cup_{g \in G} gH)(\alpha) = G(\alpha) = M$. Similarly, $\cup_{g \in G}((Hg)(\alpha)) = (\cup_{g \in G} Hg)(\alpha) = G(\alpha) = M$. Since we know by Lemma 16 that $gH \subset M$ and $Hg \subset M$ for all $g \in G$, we have that $P_L(M,N)$ and $P_R(M,N)$ are partitions of M. □

As an example, let $M = \{abc, acb, bac, bca, cab, cba\}$ and let $N = \{abc, acb\}$. Note that M and N are both GSAs. Now, let $H = \bowtie(N) = \{\mathrm{id}, (bc)\}$. Then, $H(abc) = \{abc, acb\} = N$ and $H(bac) = \{bac, cab\}$ and $H(cba) = \{cba, bca\}$. Note that

$$\{H(abc), H(bac), H(cba)\} = \{\{abc, acb\}, \{bac, cab\}, \{cba, bca\}\}$$

forms a partition of M. This partition is the right partition of M by N, denoted $P_R(M,N)$.

Also, note that $[\mathrm{id}](N) = N = \{abc, acb\}$, $[(ac)](N) = \{cba, cab\}$, and $[(ab)](N) = \{bac, bca\}$. The reader can check that $\{[\mathrm{id}](N), [(ac)](N), [(ab)](N)\}$ forms a partition of M. This partition is the left partition of M by N, denoted $P_L(M,N)$. With this notation, we now continue.

Theorem 20. *Let M and N be GSAs of a finite set S, $N \subseteq M$, and let $G = \bowtie(M)$ and $H = \bowtie(N)$. Then $H \lhd G$ if and only if $P_L(M,N) = P_R(M,N)$.*

If $H \lhd G$, we shall say that N is a ***normal subset of*** *M.*

Proof. Let M and N be GSAs of a finite set S, $N \subseteq M$, and let $G = \bowtie(M)$ and $H = \bowtie(N)$. First, we show that $H \lhd G$ implies $P_L(M,N) = P_R(M,N)$. Thus, suppose that $H \lhd G$. Then $Hg = gH$ for all $g \in G$, so that by Lemma 12, $(Hg)(\alpha) = (gH)(\alpha)$ for all $\alpha \in N$, $g \in G$. Suppose that $(Hg)(\alpha) \in P_R(M,N)$, for some $g \in G$, $\alpha \in N$; then, $(Hg)(\alpha) = (gH)(\alpha) \in P_L(M,N)$. Similarly, suppose that $(gH)(\alpha) \in P_L(M,N)$, for some $g \in G$, $\alpha \in N$; thus $(gH)(\alpha) = (Hg)(\alpha) \in P_R(M,N)$. Hence, $P_R(M,N) = P_L(M,N)$.

Next, suppose that $P_L(M,N) = P_R(M,N)$. Then for all $g \in G, \alpha \in N$, we know that $(gH)(\alpha) \in P_L(M,N)$ and hence $(gH)(\alpha) \in P_R(M,N)$. Therefore, there exists $\beta \in M$ such that $(gH)(\alpha) = H(\beta)$. Now Theorem 14 tells us that $(gH)(\alpha)$ is a GSA and that $\bowtie((gH)(\alpha)) = gHg^{-1}$. Thus, by Theorem 11, we know that $(gH)(\alpha) = (gHg^{-1})(\gamma)$ for all $\gamma \in (gH)(\alpha) = H(\beta)$. Therefore, since $\beta \in H(\beta)$ (recall that $\beta = \mathrm{id}(\beta)$, where id is the identity permutation in H), we have that $H(\beta) = (gH)(\alpha) = (gHg^{-1})(\beta)$ and since $H(\beta) = gHg^{-1}(\beta)$, Lemma 12 assures us that $H = gHg^{-1}$. Hence, since $H = gHg^{-1}$ for all $g \in G$, we have that $H \lhd G$. □

We have now translated the concept of normal subgroups into the language of arrangement sets. Now that normality is understood in terms of arrangement sets, it would seem natural to consider **quotient groups**.

As cyclic quotient groups are deeply important to the concept of solvability, our final result establishes a criterion which determines when the quotient group of the associated permutation sets of two arrangement sets is cyclic.

Theorem 21. *Let M and N be GSAs of a finite set S, $N \subseteq M$, let $G = \bowtie(M)$ and $H = \bowtie(N)$, such that $H \lhd G$, and let $\alpha \in N$. Then the quotient $\frac{G}{H}$ of G by H is cyclic if and only if there exists a permutation $f \in G$ such that there exists $n \in \mathbb{N}$ for each $T \in P_L(M,N) = P_R(M,N)$ with $T = (f^n H)(\alpha) = f^n(N)$.*

Proof. Let M and N be GSAs of a finite set S, $N \subseteq M$, and let $G = \bowtie(M)$ and $H = \bowtie(N)$, such that $H \lhd G$. Also, let $\alpha \in N$. Because $H \lhd G$, we know that $P_L(M,N) = P_R(M,N)$.

Next, suppose that $\frac{G}{H}$ is cyclic; then, there exists $f \in G$ such that $\frac{G}{H}$ is $\frac{G}{H} = \langle fH \rangle$, the cyclic group generated by fH. Thus, we know that there exists $n \in \mathbb{N}$ such that $bH = (fH)^n$ for each $bH \in \frac{G}{H}$. Also, by the definition of quotient group, $bH = (fH)^n = (f^n)H$. Consider an arbitrary arrangement set $T \in P_L(M,N) = P_R(M,N)$. We know by Lemma 17 and Theorem 15 that there exists $k \in G$ such that $T = (kH)(\alpha)$. But we already know that there exists $n \in \mathbb{N}$ such that $kH = f^n H$, whence $T = (f^n H)(\alpha)$.

To show the converse, suppose that there exists $f \in G$ such that there exists for each arrangement set $T \in P_L(M,N) = P_R(M,N)$ an $n \in \mathbb{N}$ such that $T = (f^n H)(\alpha)$. Then, for each $b \in G$, there exists $n \in \mathbb{N}$ such that $(bH)(\alpha) = (f^n H)(\alpha)$, whence we have by Lemma 12 that $bH = f^n H = (fH)^n$. Therefore, $\frac{G}{H} = \langle fH \rangle$, whereby $\frac{G}{H}$ is cyclic. □

Theorem 21 gives us all that we need to completely describe solvability in terms of GSAs. According to the modern definition of solvability, a group H_0 is **solvable** if there exists a normal chain of groups

$$H_n = \{\mathrm{id}\} \lhd H_{n-1} \lhd \cdots \lhd H_1 \lhd G = H_0,$$

where $\frac{H_i}{H_{i+1}}$ is cyclic for $0 \leq i < n$. All of the pieces of this definition now have analogs in the language of GSAs.

However, more work can be done on the relationship between GSAs and groups. A complete description of quotient groups in terms of GSAs has not been provided; it is possible that Galois's language of arrangement sets is not abstract enough to completely describe quotient groups.

Galois never provided a satisfactory definition of group in his memoir. However, it is clear that, to Galois, groups were always sets of permutations (Galois actually used the word "substitution" instead of "permutation"), and that Galois's permutation groups were always assumed to act upon arrangements. Thus, he would denote a particular group of permutations by writing the list of arrangements created when that permutation group was applied to a single arrangement.

The GSA definition provided by this paper provides a precise definition for the group concept expressed by Galois. Notice that a set of arrangements must meet only one property to be a GSA, compared to the three or four properties required by the modern group definition. The property met by GSAs is loosely related to closure, which Galois recognized was essential to his group concept. The modern group properties are immediately met by the permutation set associated with a GSA, as Theorem 10 shows.

At the end of his memoir, Galois listed out all of the arrangements of a set with four elements. He then proceeded to show that the permutation group associated with that set of arrangements is solvable; he repeatedly partitioned his list of arrangements until he had a list of arrangements that contained only one arrangement. Similarly, he showed that the permutation group associated with the set of all arrangements of five elements is not solvable, thereby showing that the general quintic polynomial is not solvable. While this fact was known prior to Galois's work, Galois was able to use his more general machinery to very quickly arrive at this result.

Galois greatly advanced in the knowledge of the theory of equations, at the same time revolutionizing modern algebra. It took others, however, to provide a satisfactory definition of the concepts Galois originated. Abstract groups today are studied in a more general context than were

Galois's permutation groups. It shall forever be true, however, that arrangement sets are key to the development of modern algebra.

Acknowledgements and Remarks

Many thanks must be given to Dr. Matt Lunsford, who had the original idea for the project and gave me a few hints along the way when they were needed. Also, the research in this paper was supported by an Undergraduate Research Grant from Union University, for which the author is very thankful.

Finally, although the precise statements of all of the definitions and theorems, as well as the proofs, are entirely original to the author, many of the results are not. Also, some notation was borrowed from [Ti], including the $\mathrm{Sym}(S)$, $H(\alpha)$, and $g(M)$ notations. The author takes responsibility for introducing the groovy bow tie notation.

Those who wish to learn more about the history of Galois theory should consult [Ed], which contains a translation of Galois's memoir. Finally, the referees of this paper must be thanked very much for many helpful comments that greatly improved the work.

References

[Ed] Harold M. Edwards: *Galois Theory.* New York: Springer-Verlag 1984.

[Ri] Laura Toti Rigatelli: *Evariste Galois.* Basel, Switzerland: Birkhauser 1996.

[Ti] Jean-Pierre Tignol: *Galois Theory of Algebraic Equations.* New Jersey: World Scientific 2001.

STUDENT ARTICLE

4

Soliton Solutions of Integrable Systems and Hirota's Method

Justin M. Curry[†]
Massachusetts Institute of Technology '08
Cambridge, MA 02139
jmcurry@mit.edu

Abstract

In this paper we investigate a general class of solutions to various partial differential equations known as solitons or stable solitary wave solutions. We introduce necessary background by considering general solutions of the classical wave equation and some of its variants, focusing on features of linearity, non-linearity, dissipation and dispersion. The Korteweg-de Vries (KdV) equation is presented as an iconic non-linear dispersive wave equation that admits soliton solutions. How soliton solutions are approximated motivates an introduction to the Padé approximation, which seeks convergence by expressing a solution as a quotient G/F of polynomials of exponentially decaying functions. The Padé approximation motivates a substitution that decouples the KdV equation into a pair of equations on the polynomials G and F. The decoupled version of the KdV equation is then greatly simplified by introducing a bilinear differentiation operator known as Hirota's D-operator. Another substitution allows Hirota's D-operator to express the KdV equation in a single bilinear form. This final form illustrates how the perturbation method can be used to produce exact soliton and multi-soliton solutions. The generation of multi-solition solutions in an almost additive fashion with this method is summarized as a non-linear superposition principle. Connections between Hirota's method, Kac-Moody algebras and quantum field theory are briefly mentioned.

4.1 Introduction

The study of the dynamical behavior of physical systems has been, and continues to be, a major source of mathematical inspiration. The twentieth century in particular has initiated a deep inquiry into a variety of non-linear systems and their unifying themes. In the spectrum of dynamics, two opposites have attracted considerable attention: chaos and solitons. Chaos theory has demonstrated that both partial and ordinary differential equations can exhibit incredibly rich behavior, allowing some deterministic systems to be exponentially unpredictable for increasing time. On the other extreme, soliton theory provides several important examples of non-linear systems behaving in a stable, quasi-linear fashion.

In this paper, we explore this second extreme and build up a concrete introduction to solitons via an inspection of the Korteweg-de Vries (KdV) equation—a non-linear dispersive equation, which is effective for describing surface waves in a shallow water domain. The existence of stable "solitary waves"—the precursors for the term "soliton"—was first discovered experimentally in 1834 by J. Scott Russell, who chased on horseback a one foot high and 30 feet long wave generated by a stopping canal boat, traveling at eight to nine miles an hour for nearly two miles in unaltered form. This solitary wave solution was re-discovered as a solution to the KdV equation in 1895 [DJ]. Since then, stable solitary wave solutions have featured prominently in many other non-linear partial differential equations (PDEs) and the methods for generating soliton solutions have led to many deep ideas in mathematics and physics.

[†] Justin Curry is a current senior in mathematics at the Massachusetts Institute of Technology. His research interests involve geometry, integrable systems, and mathematical physics. He plans on entering a PhD program in mathematics in Fall 2008.

The goal of this paper is to provide an intuition for some of these results. We begin with a simple description and definition of classical linear wave equations in Section 4.2. The one-dimensional wave equation is solved using d'Alembert's method in Section 4.2.1. Explicit plane wave solutions to the wave equation are described in Section 4.2.2 and the relevant terminology of dispersion relations, phase and group velocities are defined in Section 4.2.3. Once this relevant background is covered, we consider in Section 4.2.4 a lesser-known class of solutions to the wave equation that cannot be approximated by plane waves, called solitary waves or solitons.

Pursuant to our objective to understand which PDEs admit soliton solutions, we consider in Section 4.3 generalized wave equations that more accurately model various physical phenomena. In particular, we stress the effects of linearity, non-linearity, dispersion and dissipation on solutions to the corresponding PDEs in Sections 4.3.1, 4.3.2, 4.3.3 and 4.3.4.

Finally, we focus on the aforementioned KdV equation in Section 4.4. We first outline some of the nice properties of the KdV equation and the conservation laws it obeys in Section 4.4.1. It is stated, but not proved, that the KdV equation satisfies infinitely many conservation laws and this relates to its integrability. Our focus then returns to solitary wave solutions to the KdV equation. The fact that these solutions cannot be approximated by plane wave solutions leads us to introduce the Padé approximation in Section 4.4.2, which will motivate us to decouple the KdV equation into two equations whose solutions are polynomials of exponentially decaying functions. Padé approximation will lead us to consider in Section 4.4.3 how the perturbation method can approximate solutions to the KdV equation. Our attempt to unite Padé approximation and the perturbation method via a decoupled pair of equations will be our way of motivating and introducing Hirota's method in Section 4.4.4. A change of variables suggested by Hirota's method will allow us to put the KdV equation into a very elegant bilinear form. Before further exploring this new form, we graphically demonstrate the notion of a two-soliton solution in Section 4.4.5, and qualitatively motivate the desire to produce multi-soliton solutions to the KdV equation. The substitution suggested by the Padé approximation is then abandoned in Section 4.4.6 in favor of another change of variables for the KdV equation. This alternative bilinear form will then allow us to apply the perturbation method of Section 4.4.3 to produce exact (not approximate) multi-soliton solutions to the KdV equation in Section 4.4.7. The relationship of this powerful method to deep ideas involving Kac-Moody algebras and quantum field theory are then mentioned briefly in Section 4.5.

Special thanks must be extended to Professor Aliaa Barakat at the Massachusetts Institute of Technology for guiding the author through the rich mathematical vistas involved in integrable systems. Without her direction and support, the current paper would be non-existent.

4.2 The Wave Equation(s)

Definition 1. An equation (or system of equations) is **linear** if, whenever it has solutions u_1 and u_2, it also has $au_1 + bu_2$ as a solution, where a and b are scalar coefficients.

Definition 2. The **classical wave equation** that describes a wave propagating with constant speed c is given by the following linear partial differential equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \nabla^2 u. \tag{4.1}$$

In one dimension, equation (4.1) models the height of a plucked string as a function of space and time. More specifically, the one-dimensional wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}. \tag{4.2}$$

is an idealized model derived from using force balance and Newton's laws.

4.2.1 d'Alembert's Solution to the 1-D Wave Equation

Putting $\eta = x - ct$ and $\xi = x + ct$ we have that

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial \eta^2} - 2c^2\frac{\partial^2 u}{\partial \xi \partial \eta} + c^2\frac{\partial^2 u}{\partial \xi^2},$$
$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial \eta^2} + 2\frac{\partial^2 u}{\partial \xi \partial \eta} + \frac{\partial^2 u}{\partial \xi^2}.$$

Substituting into equation (4.2), we find

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = 0,$$

and integrating twice, we have that solutions take the form

$$u = f(\eta) + g(\xi),$$

where f and g are arbitrary functions. This corresponds to solutions propagating in the left and right directions. If we take equation (4.2) and factor accordingly

$$\left(\frac{\partial}{\partial t} - c\frac{\partial}{\partial x}\right)\left(\frac{\partial}{\partial t} + c\frac{\partial}{\partial x}\right)u = \left(\frac{\partial^2}{\partial t^2} - c^2\frac{\partial^2}{\partial x^2}\right)u = 0, \tag{4.3}$$

then d'Alembert's solution tells us that if we consider solutions to the simplified wave equation

$$u_t + cu_x = 0, \tag{4.4}$$

we are left with right-traveling solutions only, i.e.

$$u(x,t) = f(\eta) = f(x - ct).$$

4.2.2 Plane Wave Solutions

Definition 3. A **plane wave** is a solution to equation (4.1) that takes the form

$$u(\vec{x}, t) = Ae^{i(\vec{k}\cdot\vec{x} - \omega t)} \tag{4.5}$$

where i is the imaginary unit, $\vec{k}$ is the wave vector, ω is the angular frequency, and A is the (possibly complex) amplitude.

For the remainder of this paper, we will restrict our attention to one-dimensional wave equations, in which case k and x are treated as scalar-valued quantities. If x is thought of as having units of length (say meters m), then k must have units that are the corresponding inverse (m^{-1}). It is common to call k the **wave number**.

In the theory of differential equations it is common to "guess" the solution to a given equation by substituting in a function that has certain required properties (most notably it solves the provided differential equation given certain constraints). We call such an assumed form for a solution an **ansatz** for the differential equation. For example, plane waves can be taken as a good ansatz for a solution to many dierent wave equations. If we consider equation (4.4) and assume it has a solution of the form (4.5), $u(x,t) = e^{i(kx-\omega t)}$ then we find that the angular frequency and wave number must satisfy the relation

$$\omega = ck. \tag{4.6}$$

This is an example of a **dispersion relation**.

4.2.3 Dispersion Relations, Phase and Group Velocities

Definition 4. A **dispersion relation** is a relation between the energy of a system and its momentum.

Since energy in waves is proportional to frequency ω and the wave number k is proportional to momentum, equation (4.6) is an example of a dispersion relation. These sorts of relations will be very valuable when considering how fast certain Fourier components in the initial profile of a wave travel and how fast energy dissipates for the given system. One distinguishes between these notions by defining two types of velocities.

To motivate the first definition of velocity, imagine that we are watching an animation of of a one-dimensional wave equation $u(x,t)$ that can be written as function $f(kx - \omega t)$ for k and ω constant. Now imagine we are following a crest of the wave and we notice that at a specific point in time t_0 and point in space x_0, the wave has a height $H = u(x_0, t_0) = f(C_0)$ where $C_0 = kx_0 - \omega t_0$. If we allow a short amount of time Δt to elapse, we see that the point H on the curve has traveled a small distance Δx, so that $u(x_0 + \Delta x, t_0 + \Delta t) = H$. For Δt and Δx small enough, we see that this can only be the case if

$$kx_0 - \omega t_0 = C_0 = k(x_0 + \Delta x) - \omega(t_0 + \Delta t),$$

e.g. the point that gets mapped to H is the same point after the wave has traveled a small distance. This is only true if

$$k\Delta x = \omega \Delta t \qquad \Rightarrow \qquad \frac{\Delta x}{\Delta t} = \frac{\omega}{k}.$$

This is known as the **phase velocity** of the wave.

Definition 5. For a wave of the form $u(x,t) = f(kx - \omega t)$ where k and ω are constants, the **phase velocity** c_{ph} is defined as the constant

$$c_{ph} := \frac{\omega}{k}. \tag{4.7}$$

Although the phase velocity can be defined more generally, where c_{ph} determines the speed at which any one frequency component travels, we will restrict ourselves to the definition given above.

Definition 6. The propagation of energy in a system is given by the velocity of wave packets, known as the **group velocity**, which is given by

$$c_{gr} := \frac{\partial \omega}{\partial k}. \tag{4.8}$$

In the case of equation right-traveling waves $f(x - ct)$ (4.4), we determined the linear dispersion relation $\omega = ck$ (4.6). Applying the definitions for the group (4.8) and phase (4.7) velocities, we see that in this instance

$$c_{ph} = \frac{\omega}{k} = \frac{ck}{k} = c = \frac{\partial \omega}{\partial k} = c_{gr}.$$

Definition 7. A **non-dispersive wave** is a wave which is governed by a linear dispersion relation.

For linear equations, differentiation of ω gives the coefficient of k, which is similarly achieved by division. Thus a linear relation implies equality of c_{ph} and c_{gr}. Conversely, imagine that the equation $c_{ph} = c_{gr}$ holds. Treating ω as a function only of k, we can separate variables and solve the differential equation as follows:

$$\begin{aligned} \frac{d\omega}{dk} &= \frac{\omega}{k} \\ \frac{d\omega}{\omega} &= \frac{dk}{k} \\ \log \omega &= \log k + c. \end{aligned}$$

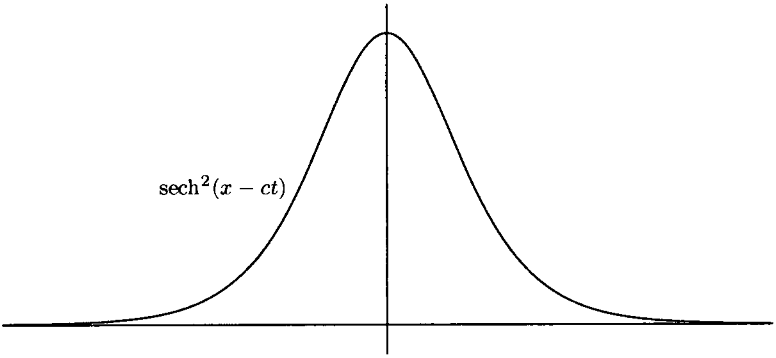


Figure 4.1: Plot of a Solitary Wave Solution

Exponentiating both sides of the equation we obtain the linear dispersion relation $\omega = Ck$ where $C = e^c$. Dispersive waves have unequal phase and group velocities, while non-dispersive waves have equal phase and group velocities.

4.2.4 Solitary Wave Solutions

Plane wave solutions are **not** the only solutions to the classical linear wave equations presented in equations (4.1), (4.2), (4.4). For example, Figure 4.1 shows a completely different solution that is not of this plane wave form. Letting

$$u(x,t) = \mathrm{sech}^2(x - ct), \tag{4.9}$$

and calculating

$$\begin{aligned} u_t &= \; - 2c\,\mathrm{sech}^2(x - ct)\tanh(x - ct), \\ u_x &= 2\,\mathrm{sech}^2(x - ct)\tanh(x - ct), \end{aligned}$$

we have thus verified that a wave of the form (4.9) satisfies our simplified right-traveling linear wave equation (4.4).

Although not obvious at this point, a solution of the form (4.9) is an example of a **solitary wave solution** or **soliton**.

Definition 8. ([DJ], [ZK]) A **solitary wave solution** or **soliton** is a solution to any wave equation that satisfies the following three properties:

1. retains its shape (initial profile) for all time,
2. is localized (asymptotically constant at $\pm\infty$ or obeys periodicity conditions imposed on the original equation),
3. can pass through other solitons and retain size and shape.

As we will see in the next few sections, other types of wave equations either permit or dismiss the possibility of solitary wave solutions. Of particular interest will be the case when the wave equation under consideration is non-linear. Certain non-linear equations will allow solitary wave solutions. In these instances, the third condition in Definition 8 will become especially important as many localized solutions tend to scatter off of one another irreversibly. This is in sharp contrast to linear equations, where two waves can pass through each other without change. Since solitons exhibit at most a phase change after interaction, we will be able to "add" (in a well-defined way to be described later) two soliton solutions to obtain a third one, achieving in effect a *non-linear superposition principle!*

4.3 Generalized Wave Equations

In the previous section, we considered an example of a **linear, non-dispersive** wave equation and the types of solutions it allows. However, these sorts of equations are often inadequate for describing the rich dynamical behavior of the universe. Physical models for vibrations, gravity waves, internal waves, surface waves and a broad range of related phenomena require a mix of dissipative, dispersive and nonlinear behavior. In this section we will consider various combinations of these features and discuss whether or not they support stable solitary wave solutions.

4.3.1 Linear Dispersive Waves

Let us consider a partial differential equation with odd spatial derivatives, such as

$$u_t + c_0 u_x + \delta u_{xxx} = 0, \tag{4.10}$$

where c_0 and δ are constant. Here we are using **subscript notation** where $u_x := \frac{\partial u}{\partial x}$, $u_{xx} := \frac{\partial^2 u}{\partial x^2}$ and so on. Taking as our ansatz the plane wave solution $u(x,t) = e^{i(kx-\omega t)}$, we get the following non-linear dispersion relation between the frequency ω and the wave number k

$$\omega = c_0 k - \delta k^3.$$

Applying the definitions for phase and group velocities we find that

$$c_0 - \delta k^2 = \frac{\omega}{k} = c_{ph} \neq c_{gr} = \frac{\partial \omega}{\partial k} = c_0 - 3\delta k^2.$$

If $\delta > 0$ we then have that

$$c_{gr} \leq c_{ph}.$$

If we assume that the Fourier transform of $u(x,0)$—call it $A(k)$—has a continuum of wave numbers in its initial profile, then the evolution of the profile is given by

$$u(x,t) = \int_{-\infty}^{\infty} A(k) e^{i(kx-\omega(k)t)} dk$$

and the Fourier components literally "disperse" or separate out according to their wave numbers. This example demonstrates that a linear dispersive wave does not exhibit stable solitary wave solutions since an initial profile consisting of many superposed wave numbers breaks apart into individual components instead of retaining its shape.

4.3.2 Linear Dissipative Waves

What is interesting to note is that we get real dispersion relations for ω whenever we have odd order derivatives for our spatial variable x. If we consider the alternative equation

$$u_t + c_0 u_x - \delta u_{xx} = 0, \tag{4.11}$$

the dispersion relation then is

$$\omega = c_0 k - i\delta k^2.$$

We thus have the solution

$$u(x,t) = e^{-k^2 t + ik(x-t)},$$

which decays exponentially with time.

Definition 9. A **dissipative wave** is a wave whose energy decreases as time increases.

In Section 4.2.3 we noted that energy in waves is proportional to frequency. This statement is true for a given amplitude, but energy is also proportional to a wave's amplitude. Although energy is a difficult concept to define generally, for a wave periodic in x with period T we may define the energy E of a wave $u(x,t)$ as

$$E := \frac{1}{2}\int_0^T |u(x,t)|^2\, dx.$$

For the above example, we find that

$$E \propto e^{-2k^2 t},$$

which clearly goes to zero as $t \to \infty$.

4.3.3 Non-Linear Non-Dispersive Waves

A common feature of equations (4.1), (4.10), and (4.11) is their linearity. With such systems, once two solutions are produced their sum is guaranteed to be a solution, and we can find a basis for the space of all solutions—the tools of linear algebra are at our disposal. Non-linear systems are much harder to study precisely for this reason. Let us consider a simple example where the wave's speed c depends on its amplitude. The equation

$$u_t + c(u)u_x = 0, \tag{4.12}$$

where $c(u) = c_0 + bu^n$ for b a constant is one such example. It is clearly non-linear, because if we consider two solutions u and v and we substitute their sum into equation (4.12), we obtain

$$u_t + v_t + c_0 + b(u+v)^n(u_x + v_x) = 0$$

if and only if $(u+v)^n = 0$, which is not true in general.

Surprisingly, the solution to equation (4.12) follows (almost) precisely d'Alembert's solution, except that c in the solution $u(x,t) = f(x-ct)$ for equation (4.4) is replaced by the function $c(u)$. This has important implications, because if $c(u)$ is increasing, then the wave travels faster as its amplitude increases, until finally the wave steepens and breaks (where "breaking" in mathematical terms is multi-valuedness).

4.3.4 Non-Linear Dispersive Waves

The upshot of the previous few subsections has been that neither linear dispersive nor non-linear non-dispersive wave equations admit solitary wave solutions. In those cases, a wave profile either has the tendency to break up according to wave number (dispersive) or steepen to multi-valuedness (non-linearity). One might imagine that a mix of these two behaviors would yield even more wild behavior, but it is surprising that these two effects can actually neutralize each other to produce soliton solutions.

We wish to provide an intuitive argument that a wave equation of the form

$$u_t + c_0 u_x + bu^n u_x + \delta u_{xxx} = 0 \tag{4.13}$$

has a solitary wave solution that retains its shape. If we suppose that equation (4.13) has a solitary wave solution as depicted in Figure 4.2, then a necessary assumption for the wave to retain its shape is the condition

$$v_{\text{top}} = v_{\text{bottom}} = v,$$

where v_{top} and v_{bottom} denote the speed at the top and bottom respectively. If we move to a coordinate system that travels with the wave we can hopefully simplify the analysis. Putting $\chi = \ell x - \Omega t$ where $v = \Omega/\ell$, so that at $\chi \sim x$ $t = 0$ and for $\chi \sim x - vt$ for all other times, we are left with a problem in terms of χ.

If $u_{\max} = A$, the amplitude of the wave, then we can approximate our solution in a neighborhood of the top as

$$u \sim A(1 - \text{const} \times \chi^2).$$

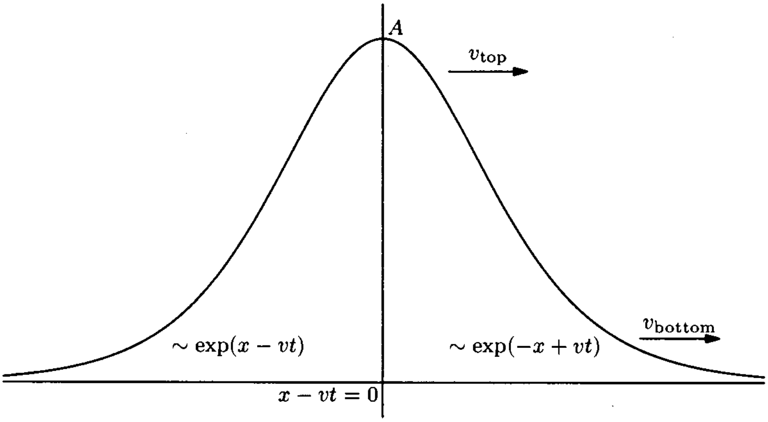


Figure 4.2: Solitary Wave Solution: Intuitive Argument

We then have that $u_{xxx} \sim 0$ in this neighborhood, so equation (4.13) reduces to exactly the equation we had in Section 4.3.3:·

$$u_t + (c_0 + bu^n)u_x \sim 0.$$

Accordingly,

$$v_{\text{top}} = c_0 + bA^n,$$

and if $b > 0$, then $v_{\text{top}} > c_0$.

At the bottom of the wave $u^n \sim 0$, and thus we can neglect the non-linear term, reducing equation (4.13) to

$$u_t + c_0 u_x + \delta u_{xxx} \sim 0. \tag{4.14}$$

Since this is exactly equation (4.10), we know that

$$c_{ph} = c_0 - \delta k^2 = v_{\text{bottom}}, \tag{4.15}$$

and if $\delta > 0$, then $v_{\text{bottom}} < c_0$. This would lead us to conclude that the velocity at the top of the wave is larger than the velocity at the bottom, and thus our wave will steepen and break. But this contradicts the stability of a solitary wave solution! Where did we go wrong in our analysis? The answer is that the expression (4.15) assumes that at the bottom u has a plane wave form! If instead we use $e^{\pm x}$ as our ansatz for (4.10), we derive the following non-linear dispersion relation

$$\Omega = c_0\ell + \delta\ell^3,$$

which in turn implies

$$v_{\text{bottom}} = \frac{\Omega}{\ell} = c_0 + \delta\ell^2.$$

We now have the important result that

$$v_{\text{top}} = v_{\text{bottom}} \Longleftrightarrow \delta\ell^2 = bA^n.$$

The above argument demonstrates several important lessons:

- A non-linear dispersive wave equation of the form (4.13) has a solitary wave solution which moves at constant speed v while preserving its shape.

- A solitary wave solution to a non-linear wave equation cannot be approximated by a plane wave solution $u \sim e^{i(kx-\omega t)}$, but rather requires an exponentially decaying solution of the form $u \sim e^{\pm(\ell x - \Omega t)}$ where $v = \Omega/\ell$.

Later in this paper we will expand on the details for the second point when we introduce a perturbation method for generating soliton solutions. To further motivate and focus our discussion we will restrict our attention to a historically important example of a non-linear dispersive wave equation, which admits soliton solutions, known as the **KdV equation**. After introducing the KdV equation and illustrating some of its properties, we will dive deeper into just how it produces soliton solutions.

4.4 The Korteweg-de Vries (KdV) Equation

One of the first PDEs for which soliton solutions were discovered is the Korteweg-de Vries (KdV) equation,

$$u_t + 6uu_x + u_{xxx} = 0.$$

As described in the introduction, this equation is useful for describing surface waves in a shallow water domain. It is straightforward to verify that

$$u(x,t) = \frac{\ell^2}{2}\operatorname{sech}^2\left(\frac{\chi}{2}\right), \qquad \chi = \ell x - \Omega t, \qquad \Omega = \ell^3, \tag{4.16}$$

is a traveling wave solution to the KdV equation.

4.4.1 Conservation Laws

One of the nice features of the KdV equation (and deeply connected to its integrability) is that it admits infinitely many conservation laws. The two bottom rungs on this conservation ladder are conservation of mass and energy. If we require the KdV equation to obey the periodicity condition

$$u(x+1,t) = u(x,t),$$

then we can prove that the "mass"

$$M := \int_0^1 u(x,t)dx$$

and the "energy"

$$E := \int_0^1 \frac{1}{2}(u(x,t))^2\, dx$$

are independent of time. Simply differentiating with respect to time we have

$$\frac{dM}{dt} = \int_0^1 u_t = \int_0^1 -6uu_x - u_{xxx} = [-3u^2]_0^1 - [u_{xx}]_0^1 = 0$$

if u and u_{xx} are assumed to be periodic in x. Using the same conditions on u, we see that

$$\begin{aligned}\frac{dE}{dt} &= \int_0^1 uu_t = -\int_0^1 6u^2u_x - \int_0^1 uu_{xxx} \\ &= -\int_0^1 \frac{\partial}{\partial x}\left(2u^3\right) - \int_0^1 \frac{\partial}{\partial x}(uu_{xx}) + \int_0^1 u_x u_{xx} = \left[\frac{1}{2}u_x^2\right]_0^1 = 0.\end{aligned}$$

Both M and E are also independent of time if we do not enforce periodicity, but rather require $u, u_x, u_{xx} \to 0$ as $x \to \pm\infty$ and where M and E are integrated on the whole real line. However, this is the case with square-integrable functions, so the demand is not too strict.

4.4.2 Padé Approximation

Although the infinity of conservation laws already points to some of the deeper aspects of the KdV equation, we are primarily concerned with how the KdV equation generates soliton solutions. As detailed in an earlier section, solitary wave solutions cannot be approximated by plane wave solutions and instead require exponentially decaying solutions of the form $e^{\pm\chi}$, where $\chi = \ell x - \Omega t$. In particular, we need to expand u in terms of $\epsilon \exp(\chi)$ where ϵ is small. Unfortunately, the right-hand side in the expression

$$u(x,t) \sim \epsilon a_1 \exp(\chi) + \epsilon^2 a_2 \exp(2\chi) + \cdots \tag{4.17}$$

may diverge for large χ, contrary to the behavior required by a solitary wave solution. Indeed, as suggested by Figure 4.2, we expect that as $\chi \to +\infty$,

$$u(x,t) \sim \exp(-\chi).$$

One way to achieve convergence is to find the Padé approximation $u = G/F$ of (4.17), where G and F are polynomials in $\exp(\chi)$.

Definition 10. Given that a function $f(x)$ is $m+n$ times differentiable, the **Padé approximant** of order (m,n) is the rational function

$$R(x) = \frac{p_0 + p_1 x + p_2 x^2 + \cdots + p_m x^m}{q_0 + q_1 x + q_2 x^2 + \cdots + q_n x^n} = \frac{G(x)}{F(x)}$$

which agrees with $f(x)$ to the highest possible order, i.e.

$$\begin{aligned} f(0) &= R(0) \\ f'(0) &= R'(0) \\ &\vdots \\ f^{(m+n)}(0) &= R^{(m+n)}(0) \end{aligned}$$

Example 11. Suppose

$$f(x) = x - x^3 + x^5 - x^7 + \cdots$$

which converges for $|x| < 1$. Factoring, we obtain an alternating geometric series in terms of x^2

$$x(1 - (x^2)^1 + (x^2)^2 - (x^2)^3 + \cdots) = \frac{x}{1+x^2} \tag{4.18}$$

and thus $f(x) \sim x^{-1}$ as $x \to \infty$. Substituting $\exp(\chi)$ for x in (4.18) gives

$$f(\exp(\chi)) = \exp(\chi) - \exp(3\chi) + \exp(5\chi) - \cdots = \frac{\exp(\chi)}{1+\exp(2\chi)} \sim \exp(-\chi)$$

as $\chi \to \infty$.

4.4.3 Perturbation Method

Considering again the KdV equation

$$u_t + 6uu_x + u_{xxx} = 0, \tag{4.19}$$

and applying the **perturbation method** one expands u as a power series in a small parameter ϵ to obtain an infinite sequence of linear equations on the components of the expansion as follows. We substitute the following expression for u in (4.19)

$$u = \epsilon u_1 + \epsilon^2 u_2 + \epsilon^3 u_3 + \cdots, \tag{4.20}$$

and by collecting like powers of ϵ we obtain the following series of equations:

$$\left(\frac{\partial}{\partial t}+\frac{\partial^3}{\partial x^3}\right)u_1=0, \tag{4.21}$$

$$\left(\frac{\partial}{\partial t}+\frac{\partial^3}{\partial x^3}\right)u_2=-6u_1\frac{\partial u_1}{\partial x}, \tag{4.22}$$

$$\left(\frac{\partial}{\partial t}+\frac{\partial^3}{\partial x^3}\right)u_3=-6\left(u_2\frac{\partial u_1}{\partial x}+u_1\frac{\partial u_2}{\partial x}\right),$$

$$\vdots$$

As explained in Section 4.13, we need to choose an exponentially decaying solution for u. Let

$$u_1=a_1\exp(\chi),\quad \chi=\ell x-\Omega t,\quad \Omega=\ell^3,$$

where a_1 and ℓ are arbitrary. Substituting this solution into (4.22), we determine that

$$u_2=a_2\exp(2\chi),\quad a_2=\frac{-a_1^2}{\ell^2}.$$

Proceeding successively, we can thus find all the u_i's in (4.20), obtaining a power series in $\epsilon\exp(\chi)$ as in (4.17). However, as already mentioned, this expression will diverge for large χ, which we can try to circumvent via a Padé approximation. The trouble with this approach is that there is no simple way to determine the required functions G and F. One potential trick is to reverse engineer a known solitary wave solution so that it takes the form G/F. In the case of (4.19) the one-soliton solution (4.16) is

$$u(x,t)=\frac{\ell^2}{2}\operatorname{sech}^2\left(\frac{\chi}{2}\right)=\frac{\ell^2}{1+\cosh(\chi)}=\frac{2\ell^2}{2+\exp(\chi)+\exp(-\chi)},$$

so

$$\frac{G}{F}=\frac{2\ell^2\exp(\chi)}{1+2\exp(\chi)+\exp(2\chi)}=\frac{2\ell^2\exp(\chi)}{(1+\exp(\chi))^2}. \tag{4.23}$$

The problem with reverse engineering is that (4.23) is an artificial byproduct of a solution we already know. What would be better is to develop a method which determines the functions G and F without a priori having the solution u. This approach is embodied by Hirota's method.

4.4.4 Hirota's Method

Our practice with the perturbation method and Padé approximations suggests that making the substitution $u(x,t)=G[\exp(\ell x-\Omega t)]/F[\exp(\ell x-\Omega t)]$ in (4.19) to obtain equations for G and F may be a fruitful undertaking. We first calculate:

$$\begin{aligned}
u&=\frac{G}{F}\\
u_t&=\frac{G_tF-GF_t}{F^2}\\
u_x&=\frac{G_xF-GF_x}{F^2}\\
u_{xxx}&=\frac{G_{xxx}}{F}-\frac{3G_{xx}F_x+3G_xF_{xx}+GF_{xxx}}{F^2}+6\frac{G_xF_x^2+GF_{xx}F_x}{F^3}-\frac{GF_x^3}{F^4}.
\end{aligned}$$

Substituting into the KdV equation (4.19), we obtain the following complicated equation

$$\begin{aligned} u_t + 6uu_x + u_{xxx} = {} & \frac{G_tF - GF_t}{F^2} + 6\frac{G}{F}\frac{G_xF - GF_x}{F^2} \\ & + \frac{G_{xxx}F - 3G_{xx}F_x - 3G_xF_{xx} - GF_{xxx}}{F^2} \\ & + 6\frac{FG_xF_x^2 + FGF_{xx}F_x - GF_x^3}{F^4} = 0 \end{aligned} \tag{4.24}$$

At first glance equation (4.24) has only made things worse. We could try to decouple this equation into a simpler set of equations. Re-expressing the $6uu_x$ term with F^3 in the denominator as one with F^4 in the denominator, we could require that individually the term with F^2 in the denominator and F^4 in the denominator are both zero:

$$G_tF - GF_t + G_{xx}F - 3G_{xx}F_x - 3G_xF_{xx} - GF_{xxx} = 0 \tag{4.25}$$

$$GG_xF^2 - G^2FF_x + G_xFF_x^2 + GFF_xF_{xx} - F_x^3G = 0. \tag{4.26}$$

Unfortunately, the G and F we derived in (4.23) do not satisfy (4.25) and (4.26), but only by a missing factor of $6G_xF_{xx}$. Changing the minus sign in front of $3G_xF_{xx}$ to a plus sign and transferring the remainder to the numerator of the F^4 term, the KdV equation (4.19) becomes

$$\begin{aligned} & \frac{G_tF - GF_t + G_{xx}F - 3G_{xx}F_x + 3G_xF_{xx} - GF_{xxx}}{F^2} + \\ & \qquad 6(G_xF - GF_x)\frac{GF - (FF_{xx} - F_x^2)}{F^4} = 0. \end{aligned}$$

Setting the terms with denominators F^2 and F^4 equal to zero, we obtain the decoupled equations:

$$G_tF - GF_t + G_{xx}F - 3G_{xx}F_x + 3G_xF_{xx} - GF_{xxx} = 0 \tag{4.27}$$

$$GF - (FF_{xx} - F_x^2) = 0. \tag{4.28}$$

We have done a great deal of work, but it doesn't appear to have paid off. However, careful analysis of the pattern of derivatives suggests that (4.27) and (4.28) can be written even more simply. Introducing a new bilinear differentiation operator, Hirota's D-operator, will greatly simplify these expressions once and for all.

Definition 12. The **Hirota D-operator** for two n-times differentiable functions f and g is defined by:

$$D_x^n f \cdot g = (\partial_{x_1} - \partial_{x_2})^n f(x_1)g(x_2)\big|_{x_1=x_2=x} \tag{4.29}$$

Example 13. We determine the following quantities

$$\begin{aligned} D_tf \cdot g &= f_tg - fg_t, \\ D_xf \cdot g &= f_xg - fg_x, \\ D_x^2f \cdot g &= f_{xx}g - 2f_xg_x + g_{xx}f, \\ D_x^2f \cdot f &= 2f_{xx}f - 2f_x^2, \\ D_x^3f \cdot g &= f_{xxx}g - 3f_{xx}g_x + 3f_xg_{xx} - fg_{xxx}. \end{aligned}$$

With the Hirota D-operator in hand, we then immediately recognize that equations (4.27) and (4.28) for G and F reduce to the following **quadratic**, also called **bilinear**, form:

$$(D_t + D_x^3)G \cdot F = 0, \tag{4.30}$$

$$2GF - D_x^2F \cdot F = 0. \tag{4.31}$$

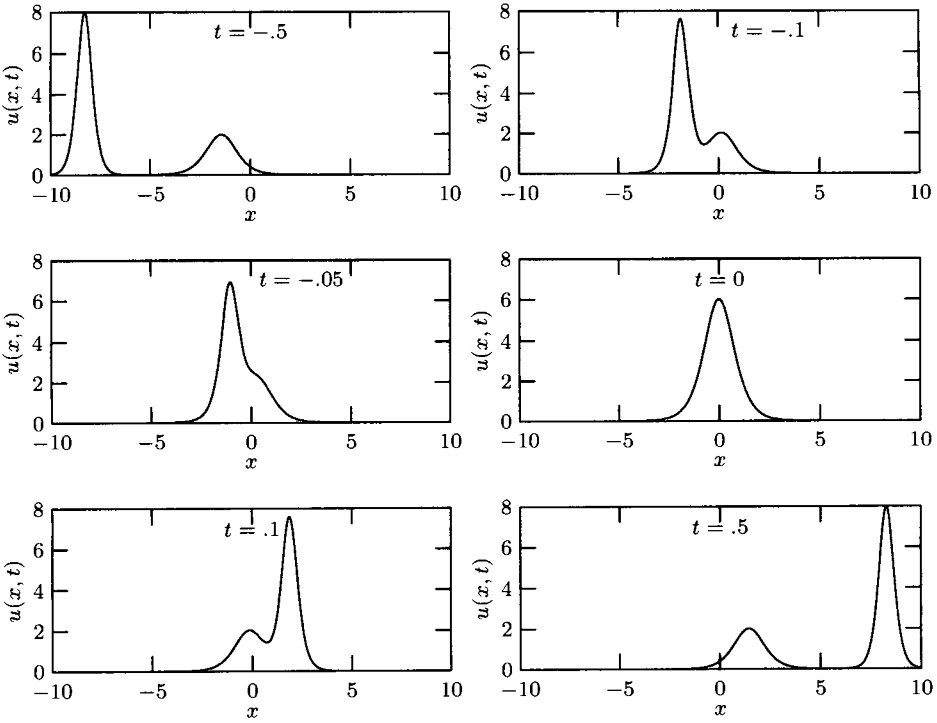


Figure 4.3: Two-Soliton Solution for Various Times

Equations (4.30) and (4.31) are the culmination of this section. Not only is the form aesthetically pleasing, but we will soon see how such a form enables one to produce soliton solutions in an almost trivial manner. Before this method for producing such solutions is presented, we would like to first understand what it means for the KdV equation to admit a **multi-soliton**, or **N-soliton**, solution.

4.4.5 N-Soliton Solutions

In Section 4.2.4 we outlined the defining characteristics of a soliton solution. In particular, we noted how the third condition in Definition 8 gives rise to a non-linear superposition principle. So far, we did nothing to illustrate this principle graphically, nor did we explain how an equation might admit multiple soliton solutions. In Figure 4.3 we have graphed the following two-soliton solution to the KdV equation (4.19) for six different times (p.75 [DJ]):

$$u(x,t) = 12\frac{3 + 4\cosh(2x - 8t) + \cosh(4x - 64t)}{[3\cosh(x - 28t) + \cosh(3x - 36t)]^2}. \tag{4.32}$$

One of the striking features of Figure 4.3 is that the tall wave actually catches up to, and passes right through, the smaller wave in an almost linear fashion. Careful inspection and exploration by the reader will reveal that after the interaction, the short wave has actually been pushed back and the tall wave has advanced forward relative to where they would have been if the waves had evolved individually, without interaction. This **phase-shift** is the trademark of the non-linearity of the KdV equation. Aside from this small difference, the ability for individual solitary waves to interact strongly and retain their shape is the defining characteristic of solitons. How multiple soliton solutions such as (4.32) are produced for the KdV equation in a more direct, algebraic fashion is the objective of the final few sections. In particular, we will find that reformulating the KdV equation into another bilinear form will allow us to simplify our analysis considerably.

4.4.6 Logarithmic Substitution for the KdV Equation

We motivated the decoupling of the KdV equation into two equations involving G and F by introducing the Padé approximant and asking for a better method to produce these polynomials in $\exp(\chi)$. We could conceivably try to express G and F in terms of a power series like (4.20), substituting into the two bilinear equations (4.30) and (4.31) and obtaining pairs of equations analogous to equations (4.21), (4.22), and so on. This, however, turns out to be a rather complicated approach as it stands, so in this section we will introduce an alternative substitution that reduces the KdV equation to another *single* equation in bilinear form. This will allow us to apply the perturbation method to produce exact multi-soliton solutions.

Instead of taking $u = G/F$, let us make the substitution

$$u = 2\frac{\partial^2}{\partial x^2} \log f.$$

Then (4.19) can be written as

$$2(\log f)_{xxt} + 3\partial_x(u^2) + u_{xxx} = 0,$$

which upon integrating once by x becomes

$$2(\log f)_{xt} + 3u^2 + u_{xx} = 0. \qquad (4.33)$$

We now calculate the relevant quantities:

$$\begin{aligned} u^2 &= 4\left(\frac{f_x}{f}\right)^4 + 4\left(\frac{f_{xx}}{f}\right)^2 - 8\frac{f_x^2 f_{xx}}{f^3}, \\ u_{xx} &= -12\left(\frac{f_x}{f}\right)^4 + 24\frac{f_x^2 f_{xx}}{f^3} - 6\left(\frac{f_{xx}}{f}\right)^2 - 8\frac{f_x f_{xxx}}{f^2} + 2\frac{f_{xxxx}}{f}, \\ 2(\log f)_{xt} &= -2\frac{f_x f_t}{f^2} + 2\frac{f_{xt}}{f}. \end{aligned}$$

After substituting into (4.33), simplifying and multiplying by $f^2/2$, we obtain the equation

$$f f_{xt} - f_x f_t + 3f_{xx}^2 - 4f_x f_{xxx} + f f_{xxxx} = 0. \qquad (4.34)$$

We are now ready to put the KdV equation into an alternate bilinear form involving the Hirota D-operator defined in equation (4.29). In example 13 we calculated several quantities using the D-operator. We now add the following two calculations:

$$\begin{aligned} D_x D_t f \cdot f &= 2(f_{tx} f - f_t f_x) && (4.35) \\ D_x^4 f \cdot f &= 2(f f_{xxxx} - 4f_x f_{xxx} + 3f_{xx}^2). && (4.36) \end{aligned}$$

If we compare equations (4.35) and (4.36) with the transformed KdV equation (4.34) we can immediately deduce the alternate bilinear form

$$D_x(D_t + D_x^3) f \cdot f = 0. \qquad (4.37)$$

Compared against (4.30) and (4.31), equation (4.37) is a simpler equation. In particular, we will find that the perturbation method produces multi-soliton solutions in a direct manner from this equation.

4.4.7 Producing N-Solitons via the D-Operator

We are now in a position to reapply the perturbation method of Section 4.4.3. It is important to note that when we try to solve PDEs such as the KdV equation via the perturbation method, we usually have an expansion of infinite order, whose coefficients we must determine successively. Truncating the solution leaves us with only approximate solutions to the original PDE. In contrast, we will find that applying the perturbation method to equations in bilinear form and choosing our early components wisely will force our infinite expansion to truncate at finite order. This will allow us to produce exact, rather than approximate, solutions via an expansion (4.20) of finite order in ϵ.

First, we take the one-soliton solution (4.16) and write for $\ell = 2$, $\Omega = \ell^3$, the solution in terms of the transformed equation for f:

$$u(x,t) = 2\operatorname{sech}^2(x-4t) = 4\frac{\partial}{\partial x}\left(\frac{e^{2x-8t}}{1+e^{2x-8t}}\right) = 2\frac{\partial^2}{\partial x^2}\log(1+e^{2x-8t}).$$

If for notational convenience we write

$$B[f,g] := D_x(D_t + D_x^3)f\cdot g,$$

we determine that for $f = 1 + e^{2x-8t}$

$$B[f,f] = B[1,1] + B[1,e^{2x-8t}] + B[e^{2x-8t},1] + B[e^{2x-8t},e^{2x-8t}] = 0,$$

which checks that this is indeed a solution to (4.37). We wish to generalize this solution to account for N-soliton solutions.

We assume that, like in Section 4.4.3, f can be expanded in positive powers of ϵ from which we can obtain an infinite sequence of equations on the components of the expansion. More precisely, we write

$$f = 1 + \sum_{n=1}^{\infty}\epsilon^n f_n(x,t),$$

and substitute this expression into (4.37), which upon collecting powers of ϵ becomes

$$B[1,1] + \epsilon(B[1,f_1] + B[f_1,1]) + \epsilon^2(B[1,f_2] + B[f_1,f_1] + B[f_2,1]) + \cdots$$
$$+\epsilon^r\left(\sum_{m=0}^{r} B[f_m,f_{r-m}]\right) + \cdots = 0. \qquad (4.38)$$

Expression (4.38) then reduces to a series of equations, where each term with common power of ϵ is required to be zero. We see, using (4.35) and (4.36), that the equation for f_1 reduces to

$$\left(\frac{\partial}{\partial t} + \frac{\partial^3}{\partial x^3}\right)f_1 = 0,$$

which we will rewrite using the following notation:

$$\hat{D} := \left(\frac{\partial}{\partial t} + \frac{\partial^3}{\partial x^3}\right) \qquad D := \hat{D}\frac{\partial}{\partial x}.$$

The first few equations from (4.38) then become

$$\hat{D}f_1 = 0 \qquad (4.39)$$
$$2Df_2 = -B[f_1,f_1] \qquad (4.40)$$
$$2Df_3 = -B[f_1,f_2] - B[f_2,f_1]. \qquad (4.41)$$

We can then easily check that, for $f_1 = \exp(\chi_1)$ where $\chi_i = \ell_i x - \ell_i^3 t + \alpha_i$ for ℓ_i and α_i arbitrary constants,

$$\hat{D}f_1 = 0, \quad B[f_1, f_1] = 0, \quad \text{and} \quad \hat{D}f_2 = 0.$$

Accordingly, we may choose $f_n = 0$ for $n = 2, 3, ...$ in expression (4.38) and we regain the solitary wave solution.

It is at this point that we make the very important observation that equation (4.39) is linear! This linearity, as we will now explain, is the key to generating multi-soliton solutions to the KdV equation. Let us assume that

$$f_1 = \exp(\chi_1) + \exp(\chi_2) \tag{4.42}$$

where χ_i is defined above. Since (4.39) is linear we know $\hat{D}f_1 = 0$. From (4.40) we have that

$$\begin{aligned} 2Df_2 = & -B[f_1, f_1] \\ = & -B[\exp(\chi_1), \exp(\chi_1)] - B[\exp(\chi_1), \exp(\chi_2)] \\ & -B[\exp(\chi_2), \exp(\chi_1)] - B[\exp(\chi_2), \exp(\chi_2)]. \end{aligned}$$

Noting the fact that only terms involving both χ_1 and χ_2 are non-zero, we find that

$$2Df_2 = -2\{(\ell_1 - \ell_2)(\ell_2^3 - \ell_1^3) + (\ell_1 - \ell_2)^4\} \exp(\chi_1 + \chi_2). \tag{4.43}$$

Equation (4.43) has a solution of the form

$$f_2 = A_2 \exp(\chi_1 + \chi_2),$$

and upon substituting into (4.43), we find that

$$A_2 = \left(\frac{\ell_1 - \ell_2}{\ell_1 + \ell_2}\right)^2.$$

Proceeding to equation (4.41), we substitute our expressions for f_2 and f_1 and determine that

$$\begin{aligned} 2Df_3 = & -A_2 B[\exp(\chi_1), \exp(\chi_1 + \chi_2)] - A_2 B[\exp(\chi_1 + \chi_2), \exp(\chi_1)] \\ & -A_2 B[\exp(\chi_2), \exp(\chi_1 + \chi_2)] - A_2 B[\exp(\chi_1 + \chi_2), \exp(\chi_2)] \\ = & -2A_2\{(-\ell_2)\ell_2^3 + (-\ell_2)^4\} \exp(2\chi_1 + \chi_2) \\ & -2A_2\{(-\ell_1)\ell_1^3 + (-\ell_1)^4\} \exp(2\chi_1 + \chi_2) \\ = & \, 0. \end{aligned}$$

Notice that in contrast to our one-soliton solution, assuming f_1 has the form (4.42) allows us to truncate (4.38) by putting $f_n = 0$ for $n \geq 3$. Putting $\epsilon = 1$, we now have an exact two-soliton solution to the KdV equation:

$$f = 1 + \exp(\chi_1) + \exp(\chi_2) + \left(\frac{\ell_1 - \ell_2}{\ell_1 + \ell_2}\right)^2 \exp(\chi_1 + \chi_2).$$

The method developed above generalizes to any exact N-soliton solution simply by putting

$$f_1 = \sum_{i=1}^{N} \exp(\chi_i) \tag{4.44}$$

and the expansion (4.38) is guaranteed to terminate after the f_N term. Although this termination can be proven, we will not do so here. It is important to note that expression (4.44) and its corresponding exact solution give us a *non-linear superposition principle.* The ability to take one-soliton solutions and combine them to form multi-soliton turns out to be an important feature of integrable systems. We could, in fact, take this as a definition of integrability.

Definition 14. ([Hi] p. 101) A set of equations written in Hirota bilinear form is **Hirota integrable**, if one can combine any number N of one-soliton solutions into an N-soliton solution.

The fact that (4.44) generates an N-soliton solution to the KdV equation (4.37) is testament to its Hirota integrability. In all cases known so far, Hirota integrability has turned out to be equivalent to more conventional definitions of integrability [Hi].

Although other methods for finding exact multi-soliton solutions exist, the Hirota D-operator is considered to be the most direct and algebraic method for doing so. The geometric and deeper connections of Hirota's method with some other areas of mathematics and physics will be discussed briefly in the next section.

4.5 Directions Forward

The field of integrable systems has seen many spectacular developments in the past several decades. This growth can largely be attributed to the fruitful exchange that occurs at the nexus of mathematics and physics—a nexus occupied by the study of Riemann surfaces, Kac-Moody algebras, twistor theory and quantum field theory. It is our hope that in this paper we have illustrated, starting with simple physical phenomenon, some of the beautiful mathematical structure that lies behind our models of the universe.

Historically speaking, Hirota's method was discovered by some of the very same brute-force calculations and by-hand manipulations carried out here [Ba]. It was only later, through the work of the Japanese mathematicians Date, Jimbo, Miwa, Kashiwara, Sato and Sato, that the deep connections between Hirota's bilinear form for non-linear PDEs and Kac-Moody algebras were discovered.

The seemingly arbitrary substitutions made to reduce the KdV equation to its bilinear form, either the rational substitution $u = G/F$ or the logarithmic substitution $u = 2(\log f)_{xx}$, are actually part of a broader class of functions known as τ-functions. The discovery that the partition functions of several important quantum field theories are τ-functions of non-linear PDEs is a major theme of current research in theoretical physics [Ba].

References

[Ba] Aliaa Barakat: Personal Discussions with the Author, 01/2008.

[DJ] P. G. Drazin and R. S. Johnson: *Solitons: an introduction*. Cambridge: Cambridge Univ. Press 1989.

[Hi] Jarmo Hietarinta: *Introduction to the Hirota Bilinear Method*, volume 638 of *Lect. Notes Phys.* New York: Springer-Verlag 2004.

[ZK] Norman J. Zabusky and Martin D. Kruskal: Interaction of "Solitons" in a Collisionless Plasma and the Recurrence of Initial States, *Phys. Rev. Lett.* **15** #6 (1965), 240–243.

FACULTY FEATURE ARTICLE

5
Tiling with Commutative Rings

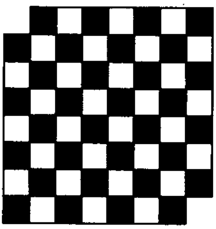
Thomas Lam[†]
Harvard University
Cambridge, MA 02138
tfylam@math.harvard.edu

Abstract

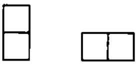
We explain an approach, due originally to Barnes, to tiling problems using some commutative algebra. We investigate in particular the occurence of coloring arguments in tiling problems. The only prerequisites are linear algebra and familiarity with rings and ideals.

5.1 A Recreational Problem

Consider the collection R of squares obtained from the chessboard by removing two opposite corners:



Can this configuration be covered with the vertical and horizontal dominoes



so that every square is covered by exactly one domino? In other words, can R be **tiled** by vertical and horizontal dominoes?

The coloring gives away the answer to this well-known problem. The region R has 32 black squares and 30 white squares. Since each domino covers exactly one black and one white square, no tiling is possible. The aim of this article is to explain a way to tackle tiling problems using a little commutative algebra. More precisely, we will explain how to obtain **coloring arguments**, similar to the above chessboard coloring, in a systematic way. I will assume that the reader is familiar with linear algebra and has seen rings and ideals before.

5.2 Tiles, Regions, and Tiling Problems

Let $\mathbb{N} = \{0, 1, 2, \ldots\}$ denote the natural numbers. A **tile** or **region** is a finite subset of $\mathbb{N}^2$ considered as a collection of boxes in the first quadrant.[1] The tiling problems that we shall consider are of the following form: given a (possibly infinite) set $\mathbf{T}$ of tiles and a region R, can R be tiled (that is, covered with tiles so that each square in R is covered once)? Each tile $\tau \in \mathbf{T}$ can be translated anywhere within $\mathbb{N}^2$ and used as many times as desired but we shall insist that *rotations and reflections*

[†] Thomas Lam was born in Hong Kong and grew up in Australia. He earned his BSc in 2001 from University of New South Wales (Australia) and received his PhD in 2005 from MIT, studying with Richard Stanley. He has been Benjamin Peirce Assistant Professor at Harvard University since then.

[1] The interested reader will have no trouble generalizing our statements to higher dimensions.

are not allowed. If we want to allow rotations of a tile then they must be added to $\mathbf{T}$ separately. Because we may translate tiles as much as we like, we will also assume that each tile $\tau \in \mathbf{T}$ has been translated as far southwest as possible, so that it touches the x- and y-axes. Thus, in the above chessboard problem, $\mathbf{T}$ consists of two elements: the vertical domino $V = \{(0,0),(0,1)\}$ and horizontal domino $H = \{(0,0),(1,0)\}$.

5.3 Coloring Arguments

Let $\mathbf{T}$ be a set of tiles. A **coloring argument** for $\mathbf{T}$ is a function $f : \mathbb{N}^2 \to \mathbb{C}$ such that

$$f(\kappa) := \sum_{(a,b)\in\kappa} f(a,b) = 0$$

for any $\kappa \subset \mathbb{N}^2$ which is a translate of a tile in $\mathbf{T}$. It is not difficult to check that the set of coloring arguments for $\mathbf{T}$ forms a vector space over $\mathbb{C}$, which we denote $\mathbb{O}(\mathbf{T})$ and shall call the **coloring space**.

If $R \subset \mathbb{N}^2$ is some region, then we say that a coloring argument $f \in \mathbb{O}(\mathbf{T})$ **forbids** R if $f(R) \neq 0$. If a coloring argument f forbids R then one immediately deduces that R is not tileable by $\mathbf{T}$. If we replace black and white by +1 and −1, then the chessboard coloring gives the following coloring argument

⋮	⋮	⋮	⋮	
−1	+1	−1	+1	⋯
+1	−1	+1	−1	⋯
−1	+1	−1	+1	⋯
+1	−1	+1	−1	⋯

(which has formula $f(a,b) = (-1)^{a+b}$) for the tile set $\mathbf{T} = \{V, H\}$ consisting of the two dominoes.

5.4 Tile Polynomials

Let us consider the polynomial ring $\mathbb{C}[x,y]$ in two variables, where $\mathbb{C}$ denotes the complex numbers. To each box $(a,b) \in \mathbb{N}^2$ in the first quadrant we associate the monomial $x^a y^b$:

⋮	⋮	⋮	⋮	
y^3	xy^3	x^2y^3	x^3y^3	⋯
y^2	xy^2	x^2y^2	x^3y^2	⋯
y	xy	x^2y	x^3y	⋯
1	x	x^2	x^3	⋯

To each region R (or tile τ) we associate the region (or tile) polynomial

$$p_R(x,y) = \sum_{(a,b)\in R} x^a y^b \in \mathbb{C}[x,y].$$

Thus, $p_V(x,y) = 1 + y$ and $p_H(x,y) = 1 + x$.

We note that translating a tile τ in the direction (a,b) corresponds to multiplying the tile polynomial by $x^a y^b$. Our assumption that the tiles $\tau \in \mathbf{T}$ are southwest-justified means that each $p_\tau(x,y)$ is not divisible by a monomial.[2]

When is a region R tileable by $\mathbf{T}$? This happens exactly when

$$p_R(x,y) = \sum_{(a,b),\tau} x^a y^b p_\tau(x,y), \tag{5.1}$$

where the summation is over some collection of translated tiles.

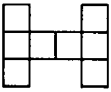
5.5 Tile Ideal

Let us define the **tile ideal** $I_\mathbf{T} \subset \mathbb{C}[x,y]$ to be the ideal generated by the tile polynomials p_τ as τ varies over the tiles in $\mathbf{T}$. A typical element of $p(x,y) \in I_\mathbf{T}$ is thus a finite linear combination

$$p(x,y) = q_1(x,y)p_{\tau_1}(x,y) + \cdots + q_k(x,y)p_{\tau_k}(x,y), \tag{5.2}$$

where $\tau_i \in \mathbf{T}$ are tiles and $q_i(x,y) \in \mathbb{C}[x,y]$. In particular, if a region R is tileable by $\mathbf{T}$ then looking at (5.1) we see that $p_R \in I_\mathbf{T}$. However, the converse is not true. The polynomials $q_i(x,y)$ in (5.2) may involve negative signs which would allow one to "remove" tiles. Let us say that a region R is **tileable by T over $\mathbb{C}$** if $p_R \in I_\mathbf{T}$. Tileability over $\mathbb{C}$ is a much easier problem, as we shall soon see.

For example, letting $R = \{(0,0),(0,1),(0,2),(1,1),(2,1),(3,0),(3,1),(3,2)\}$, we obtain:



It is easy to see that R is not tileable by the dominoes $\mathbf{T} = \{V, H\}$. However we have

$$\begin{aligned} p_R(x,y) &= 1 + y + y^2 + xy + x^2y + x^3 + x^3y + x^3y^2 \\ &= (1 + y + y^2 + x^2 + x^2y + x^2y^2 - x - xy^2)p_H(x,y) \in I_\mathbf{T}, \end{aligned}$$

so R is tileable by dominoes over $\mathbb{C}$.

5.6 Reduction to Finite Sets of Tiles

A basic theorem in commutative algebra is the **Hilbert Basis Theorem**. In our setting, it states that

Theorem 1 (Hilbert Basis Theorem). *Every ideal I in a polynomial ring $\mathbb{C}[x_1, x_2, \ldots, x_n]$ is finitely generated. Furthermore, if $S \subset I$ is any possibly infinite set of generators, then a finite subset $S' \subset S$ will generate I.*

Corollary 2. *Any possibly infinite set $\mathbf{T}$ of tiles can be replaced by a finite subset $\mathbf{T}' \subset \mathbf{T}$ of tiles, so that tileability by $\mathbf{T}$ over $\mathbb{C}$ is the same as tileability by $\mathbf{T}'$ over $\mathbb{C}$.*

Proof. Apply Theorem 1 to the tile ideal $I_\mathbf{T} \subset \mathbb{C}[x,y]$. □

5.7 Tiling Over $\mathbb{C}$ and Coloring Arguments

Proposition 3. *We have an isomorphism of $\mathbb{C}$-vector spaces*

$$\mathbb{O}(\mathbf{T}) \simeq \mathrm{Hom}_\mathbb{C}(\mathbb{C}[x,y]/I_\mathbf{T}, \mathbb{C}).$$

[2] We can avoid having to make these assumptions by using the ring $\mathbb{C}[x,y,x^{-1},y^{-1}]$ instead, but that makes other things somewhat more complicated.

Proof. Let $f \in \mathbb{O}(\mathbf{T})$. We define a $\mathbb{C}$-linear map $\phi : \mathbb{C}[x,y] \to \mathbb{C}$ by the formula

$$\phi(x^a y^b) = f(a,b)$$

and extending by linearity. Since f is a coloring argument, the map ϕ descends to a well-defined map $\bar{\phi} : \mathbb{C}[x,y]/I_{\mathbf{T}} \to \mathbb{C}$. This defines a $\mathbb{C}$-linear map $\mathbb{O}(\mathbf{T}) \to \mathrm{Hom}_{\mathbb{C}}(\mathbb{C}[x,y]/I_{\mathbf{T}}, \mathbb{C})$.

In the other direction, let $\bar{\phi} \in \mathrm{Hom}_{\mathbb{C}}(\mathbb{C}[x,y]/I_{\mathbf{T}}, \mathbb{C})$. We define $f : \mathbb{N}^2 \to \mathbb{C}$ by the formula

$$f(a,b) = \bar{\phi}(x^a y^b \bmod I_{\mathbf{T}}).$$

This f lies in $\mathbb{O}(\mathbf{T})$ and the resulting map $\mathrm{Hom}_{\mathbb{C}}(\mathbb{C}[x,y]/I_{\mathbf{T}}, \mathbb{C}) \to \mathbb{O}(\mathbf{T})$ is inverse to the one in the previous paragraph. □

It is now time for one of the main results in this article.

Theorem 4. *A region $R \subset \mathbb{N}^2$ is tileable by $\mathbf{T}$ over $\mathbb{C}$ if and only if no coloring argument $f \in \mathbb{O}(\mathbf{T})$ forbids R.*

Proof. The "only if" statement is obvious. To prove the "if" direction, we suppose that R is not tileable by $\mathbf{T}$ over $\mathbb{C}$ so that $p_R(x,y) \notin I_{\mathbf{T}}$. But this means the image $\bar{p}_R(x,y) \in \mathbb{C}[x,y]/I_{\mathbf{T}}$ is a non-zero vector in the $\mathbb{C}$-vector space $\mathbb{C}[x,y]/I_{\mathbf{T}}$. There is thus a map $\bar{\phi} \in \mathrm{Hom}_{\mathbb{C}}(\mathbb{C}[x,y]/I_{\mathbf{T}}, \mathbb{C})$ such that $\bar{\phi}(\bar{p}_R) \neq 0$. Using the isomorphism of Proposition 3 this gives a coloring argument $f \in \mathbb{O}(\mathbf{T})$ such that $f(R) \neq 0$. □

5.8 Nullstellensatz and Varieties

Let $I \subset \mathbb{C}[x,y]$ be an ideal. We define the **variety** $V(I)$ of I to be

$$V(I) = \{(\alpha,\beta) \in \mathbb{C}^2 \mid p(\alpha,\beta) = 0 \text{ for every } p(x,y) \in I\}.$$

If $X \subset \mathbb{C}^2$ is a set of points in the plane we define the **ideal** $I(X) \subset \mathbb{C}[x,y]$ **of** X by

$$I(X) = \{p(x,y) \in \mathbb{C}[x,y] \mid p(\alpha,\beta) = 0 \text{ for every } (\alpha,\beta) \in X\}.$$

(One can obviously make these definitions in dimensions more than two.)

An ideal I in a commutative ring B is called **radical** if for any $b \in B$ such that $b^n \in I$ we have $b \in I$. For example, the ideal $\langle 1+x, 1+y\rangle \subset \mathbb{C}[x,y]$ that we have previously seen is radical. A fundamental result in commutative algebra and algebraic geometry is **Hilbert's Nullstellensatz**.

Theorem 5 (Nullstellensatz). *Let $I \subset \mathbb{C}[x_1, x_2, \ldots, x_n]$ be an ideal not equal to the whole polynomial ring. Then $V(I)$ is non-empty. Furthermore, if I is radical then we have $I(V(I)) = I$.*

5.9 Tile Variety

Theorem 4 is satisfying theoretically, but to solve our favorite tiling problems it would be nice to exhibit an explicit basis for $\mathbb{O}(\mathbf{T})$. By Proposition 3, the dimension of $\mathbb{O}(\mathbf{T})$ is equal to that of $\mathrm{Hom}_{\mathbb{C}}(\mathbb{C}[x,y]/I_{\mathbf{T}}, \mathbb{C})$. If $\mathbb{C}[x,y]/I_{\mathbf{T}}$ is infinite-dimensional over $\mathbb{C}$ (it will always be of countable dimension), then $\mathrm{Hom}_{\mathbb{C}}(\mathbb{C}[x,y]/I_{\mathbf{T}}, \mathbb{C})$ will be of uncountable dimension. As an example, take $\mathbf{T} = \{V\}$ to consist of only the vertical domino. Then $\mathbb{C}[x,y]/I_{\mathbf{T}} \simeq \mathbb{C}[x]$ is infinite-dimensional over $\mathbb{C}$. For simplicity we will assume that $\mathbb{C}[x,y]/I_{\mathbf{T}}$ and thus $\mathbb{O}(\mathbf{T})$ is a finite-dimensional $\mathbb{C}$-vector space.[3]

Define the tile variety $V_{\mathbf{T}} = V(I_{\mathbf{T}}) \subset \mathbb{C}^2$ to be the variety associated to the ideal $I_{\mathbf{T}}$. For example, if $\mathbf{T} = \{V, H\}$ then $V_{\mathbf{T}}$ is given by the set of common zeroes of $1+x$ and $1+y$. Thus $V_{\mathbf{T}} = \{(-1,-1)\}$. It will follow from Theorem 6 below that if $\mathbb{C}[x,y]/I_{\mathbf{T}}$ is finite-dimensional over $\mathbb{C}$ then $V_{\mathbf{T}}$ is a finite set of points.

[3] The description we now give will not lead to a basis for $\mathbb{O}(\mathbf{T})$ in the infinite-dimensional case, but other techniques such as **Gröbner bases** can still tackle the general case.

For a point $(\alpha, \beta) \in V_{\mathbf{T}}$ define a map $\bar{\phi}_{\alpha,\beta} \in \mathrm{Hom}_{\mathbb{C}}(\mathbb{C}[x,y]/I_{\mathbf{T}}, \mathbb{C})$ by evaluating polynomials at (α, β):

$$\bar{\phi}_{\alpha,\beta}(p(x,y)) = p(\alpha,\beta).$$

Note that this equations is well-defined exactly because $(\alpha,\beta) \in V_{\mathbf{T}}$. These elements of $\mathrm{Hom}_{\mathbb{C}}(\mathbb{C}[x,y]/I_{\mathbf{T}}, \mathbb{C})$ are very special: they are not just linear maps, but also $\mathbb{C}$-algebra homomorphisms of $\mathbb{C}[x,y]/I_{\mathbf{T}}$ to $\mathbb{C}$. Under the isomorphism of Proposition 3, $\bar{\phi}_{\alpha,\beta}$ corresponds to the coloring argument $f : \mathbb{N}^2 \to \mathbb{C}$ given by $f_{\alpha,\beta}(a,b) = \alpha^a \beta^b$.

Perhaps you now see where we are heading. If we take $\mathbf{T} = \{V, H\}$ to consist of the two dominoes, and $(\alpha,\beta) = (-1,-1)$ then $f_{-1,-1}(a,b) = (-1)^{a+b}$ is just the black-white chessboard coloring!

5.10 A Basis for the Coloring Space

Theorem 6. *Suppose $\mathbb{C}[x,y]/I_{\mathbf{T}}$ has dimension n over $\mathbb{C}$ and $I_{\mathbf{T}}$ is a radical ideal. Then $V_{\mathbf{T}} = \{(\alpha_1,\beta_1),\ldots,(\alpha_n,\beta_n)\}$ consists of n points and the set $\{f_{\alpha_i,\beta_i} \in \mathbb{O}(\mathbf{T})\}$ forms a basis of the coloring space $\mathbb{O}(\mathbf{T})$.*

Proof. We claim that an element $\bar{p}(x,y) \in \mathbb{C}[x,y]/I_{\mathbf{T}}$ is completely determined by its values $\bar{p}(\alpha_i,\beta_i)$ on $V_{\mathbf{T}}$. This follows from Theorem 5: if $p, q \in \mathbb{C}[x,y]$ take the same values everywhere on $V_{\mathbf{T}}$ then the difference $p - q$ lies in $I(V_{\mathbf{T}})$ and thus in $I_{\mathbf{T}}$ by the Nullstellensatz. In particular, we have

$$\dim_{\mathbb{C}}(\mathbb{C}[x,y]/I_{\mathbf{T}}) \le |V_{\mathbf{T}}|.$$

But if $\{(\alpha_1,\beta_1),\ldots,(\alpha_m,\beta_m)\} \subset V_{\mathbf{T}}$ and $j \in [1,m]$ is fixed let us pick for each $i \neq j$ in $[1,m]$ a polynomial

$$q_i^{(j)}(x,y) = \frac{x-\alpha_i}{\alpha_j-\alpha_i} \quad \text{or} \quad q_i^{(j)}(x,y) = \frac{y-\beta_i}{\beta_j-\beta_i},$$

insisting that we choose an expression such that the denominator is non-zero (most of the time either one will do). Then the product

$$q^{(j)}(x,y) = \prod_{i\neq j} q_i^{(j)}(x,y) \in \mathbb{C}[x,y]$$

takes the value 1 at (α_j,β_j) and the value 0 at every other (α_i,β_i). These m polynomials give m linearly independent elements of $\mathbb{C}[x,y]/I_{\mathbf{T}}$. Thus,

$$\dim_{\mathbb{C}}(\mathbb{C}[x,y]/I_{\mathbf{T}}) \ge |V_{\mathbf{T}}|$$

and we conclude that $n = \dim_{\mathbb{C}}(\mathbb{C}[x,y]/I_{\mathbf{T}}) = |V_{\mathbf{T}}|$. In particular, we have shown that $V_{\mathbf{T}} = \{(\alpha_1,\beta_1),\ldots,(\alpha_n,\beta_n)\}$ is finite. One checks that the maps $\{\bar{\phi}_{\alpha_i,\beta_i}\} \subset \mathrm{Hom}_{\mathbb{C}}(\mathbb{C}[x,y]/I_{\mathbf{T}}, \mathbb{C})$ form a dual-basis to $\{q^{(j)}(x,y)\} \subset \mathbb{C}[x,y]/I_{\mathbf{T}}$, to complete the proof. □

For $\mathbf{T} = \{V, H\}$, we have remarked that $I_{\mathbf{T}}$ is radical so Theorem 6 says that the chessboard coloring is essentially the only coloring argument. There is also a version of Theorem 6 which applies even when $I_{\mathbf{T}}$ is not radical.

5.11 Summary of Strategy

Let us summarize our approach to a tiling problem. We are given a set $\mathbf{T}$ of tiles and a region R. First, we convert each tile $\tau \in \mathbf{T}$ into a polynomial $p_\tau(x,y)$. We (try to) solve all these polynomials simultaneously, to find the tile variety $V_{\mathbf{T}} \subset \mathbb{C}^2$. If $V_{\mathbf{T}} = \varnothing$ then every region R is tileable by $\mathbf{T}$ over $\mathbb{C}$.

We suppose $V_{\mathbf{T}}$ consists of a finite set of points. Next we evaluate $p_R(x,y)$ at each point (α,β) of $V_{\mathbf{T}}$. If for some point we have $p_R(\alpha,\beta) \neq 0$ then we have found a coloring argument $f_{\alpha,\beta}$ which forbids R. If not, but in addition we know that $I_{\mathbf{T}}$ is radical, then we can conclude from Theorem 6 that no coloring argument can show that R is not tileable. Of course, to completely resolve whether R is tileable by $\mathbf{T}$ is a much harder problem.

Furthermore, all the results so far work in any number of dimensions.

5.12 Final Comments

Essentially all of what we have presented so far is a simplification of work of Barnes [B1, B2]. However, much more can be said if we are willing to restrict our class of tiling problems. Let us now assume that all the tiles and regions that we consider are **bricks**. In two-dimensions, bricks are just rectangles. In d-dimensions, they are regions of the form $[a_1, b_1] \times \cdots \times [a_d, b_d]$.

A fundamental result is an analogue of the Hilbert Basis Theorem over $\mathbb{N}$, due to de Bruijn and Klarner.

Theorem 7 ([dBK]). *When considering tiling problems of bricks by bricks, any collection of brick tiles can be replaced by a finite subcollection.*

For brick tiling problems, tiling over $\mathbb{C}$ and usual tilings are not too different. Barnes proved:

Theorem 8 ([B2]). *Let* $\mathbf{T}$ *be a finite set of brick tiles. Then there is some constant* K *such that every brick region* R *with all dimensions greater than* K *can be tiled by* $\mathbf{T}$ *if and only if it can be tiled by* $\mathbf{T}$ *over* $\mathbb{C}$.

Together with Ezra Miller and Igor Pak, I have been studying some computational issues for tilings. I now describe some of our results. Let us say that a set $\mathbf{S}$ of bricks has a **finite description** if it is a finite union $\mathbf{S} = \cup_i \mathbf{S}_i$ of brick classes $\mathbf{S}_i$ such that each class is of one of the following forms:

1. $\{(l_1, \ldots, l_d) \mid l_i = a\}$

2. $\{(l_1, \ldots, l_d) \mid l_i > a\}$

3. $\{(l_1, \ldots, l_d) \mid l_i > a$ and $l_i \equiv b \bmod c$

for integers a, b and c.

Proposition 9 ([LMP]). *Let* $\mathbf{T}$ *be a set of bricks. Then the set* $\mathbf{S}$ *of bricks which can be tiled by* $\mathbf{T}$ *admits a finite description.*

Theorem 10 ([LMP]). *Suppose we are in* $d = 2$ *dimensions and* $\mathbf{T}$ *is a finite set of bricks. Then it is possible to compute a finite description for the set* $\mathbf{S}$ *of bricks tileable by* $\mathbf{T}$.

Surprisingly, we conjecture that Theorem 10 fails in higher dimensions. That is, when $d \geq 3$, a finite description for the set $\mathbf{S}$ of bricks tileable by $\mathbf{T}$ is not computable.

References

[B1] Frank W. Barnes: Algebraic Theory of Brick Packing I, *Discrete Math.* **42** (1982), 7–26.

[B2] Frank W. Barnes: Algebraic Theory of Brick Packing II, *Discrete Math.* **42** (1982), 129–144.

[dBK] Nicolaas G. de Bruijn and David A. Klarner: A finite basis theorem for packing boxes with bricks, *Phillip Res. Repts* **30** (1975) 337*–343*.

[LMP] Thomas Lam, Ezra Miller, and Igor Pak: Brick tilings, in preparation.

FACULTY FEATURE ARTICLE

6 Twisting with Fibonacci

Dana Rowland[†]
Merrimack College
North Andover, MA 01845
Dana.Rowland@merrimack.edu

Abstract

Determining when two links are equivalent is one of the central goals of knot theory. This paper describes the Conway polynomial, a link invariant that offers one approach to this problem. When calculating the Conway polynomial of the $(n, 2)$ torus knots, we encounter the familiar patterns of Pascal's triangle and the Fibonacci sequence.

6.1 Introduction

Pick up a piece of string. Tangle it up, twist it around, knot it up, and then attach the ends. The result is a mathematical knot. Suppose a friend does the same thing. Try to twist, stretch, or otherwise deform the two tangled loops, without cutting the strings, so that they look exactly the same. If this is possible, then the knots are said to be equivalent. Determining whether or not two knots are equivalent is one of the central questions in knot theory.

Mathematically, a **knot** is a continuous closed curve in space that does not intersect itself. A **link** is the disjoint union of finitely many knots, where the number of components of the link is determined by the number of knots. If each component is assigned a direction, then the link is **oriented**. Two oriented links are equivalent if one can be deformed into the other in such a way that the orientation is preserved. A two-dimensional picture of a link is called a **projection.**

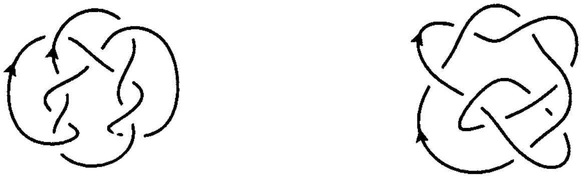


Figure 6.1: Do these two oriented link projections represent equivalent links?

Knot theorists use **link invariants** to distinguish between different links. If two links are equivalent, then calculating the link invariant for each link produces the same result, despite the fact that the link projections may appear to be drastically different.

One example of a link invariant is the **Conway polynomial**. Given a projection of an oriented link L, we can assign a polynomial $\nabla(L)$, described using the variable z. The polynomial is defined

[†] Dana Rowland has been teaching in the mathematics department at Merrimack College since 2001, where she is now an associate professor. She earned her doctorate in mathematics in 2001 at Stanford University under Ralph Cohen, in the area of algebraic topology. As an undergraduate, she majored in both mathematics and English at the University of Notre Dame. Her current research interests include knot theory and graph theory. When she is not doing mathematics, Dr. Rowland enjoys playing bassoon and valve trombone in the Merrimack College jazz ensemble and being constantly surprised by her amazing children Ben (4 years) and Michela (21 months).

so that no matter how the link is twisted around in space, the polynomial will not change—any two projections of the same link will have the same Conway polynomial.

The Conway polynomial of a particular projection is calculated recursively by applying the following two definitions.

Definition 1. If L is equivalent to a single unknotted circle ("the unknot"), then its Conway polynomial is equal to 1; that is,

$$\nabla\left(\bigcirc\right) = 1. \qquad (6.1)$$

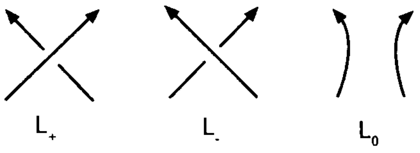


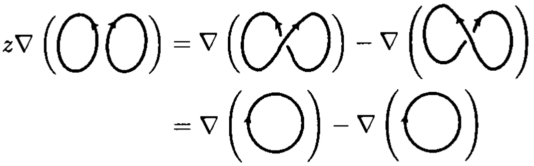
Figure 6.2: These three link projections are identical outside of the region shown.

Definition 2. Suppose L_+, L_-, and L_0 are three oriented link projections that are identical except near one crossing of L_+, where they appear as in Figure 6.2. Then their Conway polynomials satisfy the relation

$$z\nabla(L_0) = \nabla(L_+) - \nabla(L_-). \qquad (6.2)$$

These two defintions, together with the requirement that two projections of the same link must have the same Conway polynomial, suffice for calculating the Conway polynomial of any knot or link. We can always unknot a link by changing finitely many crossings. By applying Definition 2 at one of these crossings and appropriately assigning the roles of L_+, L_-, and L_0, we can find the Conway polynomial of a given link in terms of the Conway polynomials of links with fewer crossings. We can eliminate all crossings by repeating this process, resulting in only the trivial knot or trivial links.

By Definition 1, we know $\nabla\left(\bigcirc\right) = 1$. The Conway polynomial of a trivial link is 0. To see this, label the trivial link as L_0 and form L_+ and L_- by joining two of the components using a positive and negative crossing respectively. Applying (6.2) gives



$$= 0.$$

The following example illustrates how these two definitions are used to calculate the Conway polynomial of the trefoil knot.

Example 3. The Conway polynomial of the trefoil knot is $1 + z^2$.

Proof. Select one crossing in the trefoil knot. Since the crossing is a positive crossing, label the trefoil knot as L_+ and replace the region near the chosen crossing as shown in Figure 6.2 to obtain

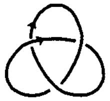
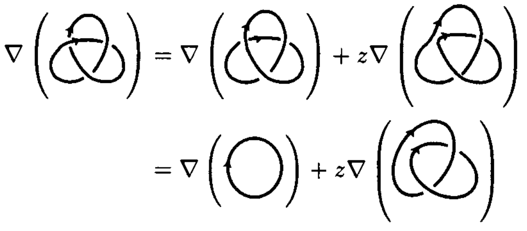
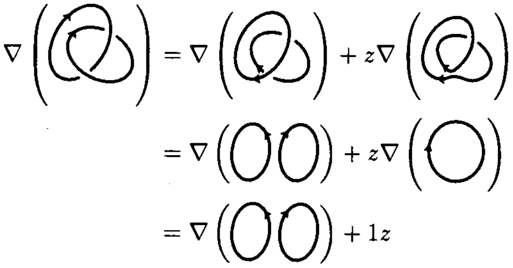


Figure 6.3: The trefoil knot.

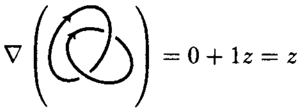
L_- and L_0. Then use (6.2).



The Conway polynomial of the unknot is 1. To find the Conway polynomial of the link, select another crossing and apply (6.2) again.



Since the Conway polynomial of the trivial link is 0, we see that



and the Conway polynomial of the trefoil is $1 + z(z) = 1 + z^2$, as claimed. □

Before continuing, we urge the reader to try a problem or two.

Exercise 4. A different projection of the trefoil knot is shown in Figure 6.4. Make this knot out of string and manipulate the string to show that this knot is equivalent to the one used in Example 3. Verify that calculating the Conway polynomial starting with this projection also results in $1 + z^2$.

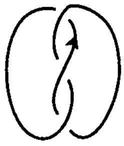


Figure 6.4: Another projection of the trefoil knot

Exercise 5. Show that the Conway polynomial of the figure-eight knot, shown in Figure 6.5, is $1 - z^2$.

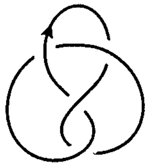


Figure 6.5: The figure-eight knot.

6.2 Torus Links

Suppose you have two strings, side by side, and the tops of the strings are fixed. Twist the right string over the left string n times, and orient both strings in the same direction. Without introducing any additional crossings, attach the bottoms of the strings to the tops, as shown in Figure 6.6. The resulting knot or link is known as an $(n, 2)$ torus link, because it lies flat on a torus, wrapping around twice longitudinally and n times through the center.[1]

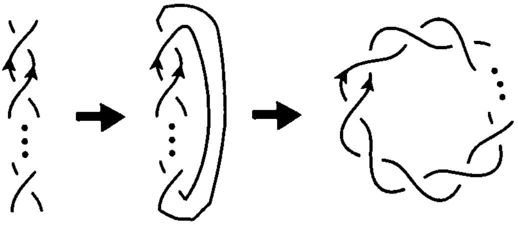


Figure 6.6: An $(n, 2)$ torus link.

Now we explore what happens when we calculate the Conway polynomials of these links. Let $\mathcal{L}_n$ denote the $(n, 2)$ torus link. Notice that $\mathcal{L}_1$ is the unknot, $\mathcal{L}_2$ is the link from Example 3, and $\mathcal{L}_3$ is the trefoil. When n is odd, $\mathcal{L}_n$ is a knot. When n is even, $\mathcal{L}_n$ is a two-component link. Furthermore, for $n \geq 3$, changing a single crossing in the $(n, 2)$ torus link results in a projection of the $(n - 2, 2)$ torus link. See Figure 6.7.

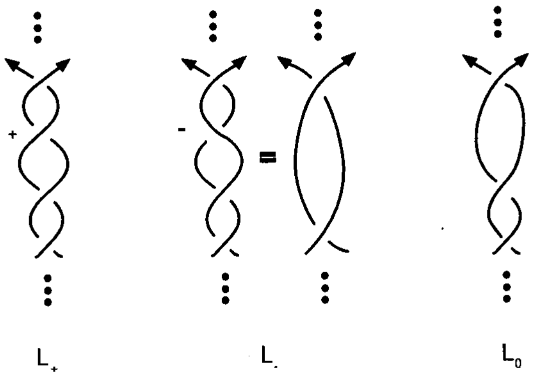


Figure 6.7: Applying (6.2) to torus links with $L_+ = \mathcal{L}_n$, $L_- = \mathcal{L}_{n-2}$, and $L_0 = \mathcal{L}_{n-1}$.

[1] See [Ad, Section 5.1] for a general description of (n, m) torus knots and links.

Using the definitions for calculating the Conway polynomial, we observe that for all $n \geq 3$,

$$\nabla(\mathcal{L}_n) = \nabla(\mathcal{L}_{n-2}) + z\nabla(\mathcal{L}_{n-1}). \tag{6.3}$$

This quickly leads to a table of Conway polynomials:

$$\begin{aligned}
\nabla(\mathcal{L}_1) &= 1 \\
\nabla(\mathcal{L}_2) &= z \\
\nabla(\mathcal{L}_3) &= 1 + z^2 \\
\nabla(\mathcal{L}_4) &= 2z + z^3 \\
\nabla(\mathcal{L}_5) &= 1 + 3z^2 + z^4 \\
\nabla(\mathcal{L}_6) &= 3z + 4z^3 + z^5 \\
\nabla(\mathcal{L}_7) &= 1 + 6z^2 + 5z^4 + z^6 \\
\nabla(\mathcal{L}_8) &= 4z + 10z^3 + 6z^5 + z^7 \\
\nabla(\mathcal{L}_9) &= 1 + 10z^2 + 15z^4 + 7z^6 + z^8 \\
\vdots \quad & \qquad \vdots
\end{aligned}$$

6.3 Pattern Recognition and Formulas

If we look at the table of Conway polynomials, we can immediately make a few observations. First, the degree of the polynomial for the $(n, 2)$ torus link is $n - 1$. The polynomials are all **monic**, meaning that the highest order term has 1 as its coefficient. When n is odd, the constant term of the polynomial is 1 and the polynomial contains only even powers of z. When n is even, the polynomial contains only odd powers of z, and the coefficient of the z term is $n/2$.

These observations begin to reveal the behavior of the Conway polynomial of the $(n, 2)$ torus link, but we can do better by taking a closer look at the pattern formed by all the coefficients:

$$\begin{array}{lccccccccc}
\nabla(\mathcal{L}_1) = & 1 \\
\nabla(\mathcal{L}_2) = & & 1z \\
\nabla(\mathcal{L}_3) = & 1 & + & 1z^2 \\
\nabla(\mathcal{L}_4) = & & 2z & + & 1z^3 \\
\nabla(\mathcal{L}_5) = & 1 & + & 3z^2 & + & 1z^4 \\
\nabla(\mathcal{L}_6) = & & 3z & + & 4z^3 & + & 1z^5 \\
\nabla(\mathcal{L}_7) = & 1 & + & 6z^2 & + & 5z^4 & + & 1z^6 \\
\nabla(\mathcal{L}_8) = & & 4z & + & 10z^3 & + & 6z^5 & + & 1z^7 \\
\nabla(\mathcal{L}_9) = & 1 & + & 10z^2 & + & 15z^4 & + & 7z^6 & + & 1z^8
\end{array}$$

Notice the appearance of Pascal's triangle along the diagonals of the Conway polynomial coefficients! Alternatively, we can find these coefficients of the Conway polynomials within Pascal's triangle, as seen in Figure 6.8.

Recall that the entries in Pascal's triangle are the binomial coefficients. This suggests the following formulae.

Theorem 6. *The Conway polynomial of the* $(2n + 1, 2)$ *torus knots is given by the equation*

$$\begin{aligned}
\nabla(\mathcal{L}_{2n+1}) &= \binom{n}{0} + \binom{n+1}{2}z^2 + \binom{n+2}{4}z^4 + \cdots + \binom{2n}{2n}z^{2n} \\
&= \sum_{j=0}^{n} \binom{n+j}{2j} z^{2j} \qquad (6.4)
\end{aligned}$$

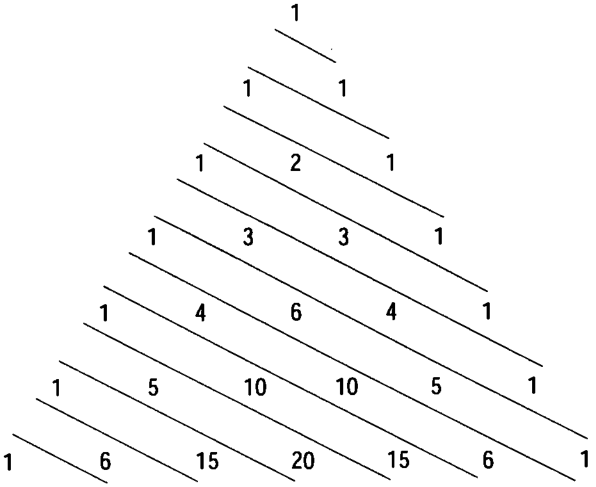


Figure 6.8: The coefficients of the Conway polynomials of the $(n, 2)$ torus knots and links can be found within Pascal's triangle.

and the Conway polynomial of the $(2n, 2)$ *torus links is given by*

$$\begin{aligned}\nabla(\mathcal{L}_{2n}) &= \binom{n}{1}z + \binom{n+1}{3}z^3 + \binom{n+2}{5}z^5 + \cdots + \binom{2n-1}{2n-1}z^{2n-1} \\ &= \sum_{j=0}^{n-1}\binom{n+j}{2j+1}z^{2j+1}. \qquad (6.5)\end{aligned}$$

Proof. We prove (6.4) and (6.5) using the principle of mathematical induction. Since $\mathcal{L}_1$ is the unknot, we know that $\nabla(\mathcal{L}_1) = 1 = \binom{0}{0}$. In Example 3, we showed that $\nabla(\mathcal{L}_2) = z = \binom{0}{0}z$, so the results are valid when $n = 0$.

Assume we know the formulae hold for all positive integers less than m. We prove the formula holds when $m = 2n$ is even. (The case for m odd is similar, and is left to the reader.) We have

$$\begin{aligned}\nabla(\mathcal{L}_{2n}) &= \nabla(\mathcal{L}_{2n-2}) + z\nabla(\mathcal{L}_{2n-1}) \\ &= \sum_{j=0}^{n-2}\binom{n-1+j}{2j+1}z^{2j+1} + z\sum_{j=0}^{n-1}\binom{n-1+j}{2j}z^{2j} \\ &= \sum_{j=0}^{n-2}\left(\binom{n-1+j}{2j+1} + \binom{n-1+j}{2j}\right)z^{2j+1} + z^{2n-1} \\ &= \sum_{j=0}^{n-1}\binom{n+j}{2j+1}z^{2j+1}. \qquad \square\end{aligned}$$

The sudden appearance of Pascal's triangle allowed us to conjecture and prove a result about the Conway polynomials of mathematical knots. As frequently occurs in mathematics, results from a seemingly unrelated field can be utilized where least expected.

6.4 Torus Links and the Fibonacci Sequence

In [Ka], Kauffman observed another relationship: If we evaluate the Conway polynomials of these torus links at $z = 1$, then we obtain the Fibonacci sequence.

Recall that the Fibonacci sequence $\{f_n\}$ is defined by the following recursive relation.

$$\begin{aligned} f_1 &= 1 \\ f_2 &= 1 \\ f_n &= f_{n-2} + f_{n-1} \end{aligned}$$

When $z = 1$, we see that $\nabla(\mathcal{L}_1)|_{z=1} = 1$, $\nabla(\mathcal{L}_2)|_{z=1} = 1$, and the recursive relation from (6.3) becomes

$$\nabla(\mathcal{L}_n)\big|_{z=1} = \nabla(\mathcal{L}_{n-2})\big|_{z=1} + 1 \cdot \nabla(\mathcal{L}_{n-1})\big|_{z=1},$$

which establishes the identity

$$\nabla(\mathcal{L}_n)\big|_{z=1} = f_n.$$

Combining this with (6.4) and (6.5), we obtain the identities

$$f_{2n+1} = \sum_{j=0}^{n} \binom{n+j}{2j} \tag{6.6}$$

$$f_{2n} = \sum_{j=0}^{n-1} \binom{n+j}{2j+1}. \tag{6.7}$$

The fact that the Fibonacci numbers occur as sums of the diagonals of Pascal's triangle shown in Figure 6.8 was discovered by Edouard Lucas in 1876. See [Ko] for a direct proof.

6.5 Conclusion

Since the polynomials given in (6.4) and (6.5) are distinct, we know that the $(n, 2)$ torus knots and links are distinct for different values of n. The Conway polynomial can be a useful tool for telling many knots and links apart, including those in Figure 6.1.

Exercise 7. Use the Conway polynomial to prove that the oriented link projections in Figure 6.1 do *not* represent equivalent links.

The Conway polynomial is equivalent to a normalized version of the very first polynomial for knots and links, which was invented by J. Alexander in 1928. Alexander defined his original polynomial using the **Seifert matrix** constructed from a surface spanning an oriented link, and in 1969 John Conway discovered the polynomial could be derived more simply using the above definitions. The Alexander-Conway polynomial was one of the major tools used for distinguishing knots for the next 60 years.

However, the Conway polynomial cannot distinguish between all knots or links. Two knots or links can have the same Conway polynomials even if they are not equivalent. For example, **splittable links** are links that can be deformed so that the components lie on different sides of a plane in $\mathbb{R}^3$.

Exercise 8. Show that the Conway polynomial of any splittable link is 0. (Hint: Label the splittable link as L_0.)

Another example which illustrates the limitations of the Conway polynomial is the 11-crossing Conway knot, shown in Figure 6.9. This knot has Conway polynomial 1, even though it cannot be unknotted.

In fact, given any knot, there are infinitely many other knots with the same Conway polynomial (*cf.* [Cr, p. 164]).

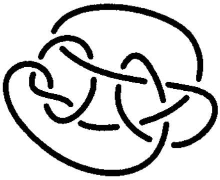


Figure 6.9: This 11 crossing knot has the same Conway polynomial as the unknot.

In 1984, Vaughan Jones discovered a connection between **von Neumann algebras** and **braid groups**, which led to a new polynomial for knots and links. The **Jones polynomial** has the advantage that there are no known examples of knots that have the same Jones polynomial as the trivial knot. Jones' discovery encouraged other mathematicians to search for more knot polynomials. This quickly led to the discovery of the **Homfly polynomial**, a two-variable generalization of both the Alexander-Conway and Jones polynomials which was developed independently by Jim Hoste, Adrian Ocneanu, Raymond Lickorish and Ken Millett, Peter Freyd and David Yetter, and Jozef Przytycki and Pawel Traczyk. Still other polynomial invariants grew out of a surprising connection between knot theory and theoretical physics.

The connections between knot polynomials and previously unrelated fields of mathematics led to renewed interest in the mathematical theory of knots and links, and rapid advances in the subject. Still, none of these polynomials provides a complete invariant—infinitely many examples exist of pairs of non-equivalent knots that still have the same knot polynomials. Knot theorists continue to search for more ways to distinguish knots and links.

We refer the interested reader to [Ad, Cr, Sc] for more about knot theory and to [BQ, En] for additional properties of Pascal's triangle and the Fibonacci sequence. We conclude with a final problem.

Take a rubber band, and twist it n times while holding onto two ends of the rubber band. Link the ends together using a clasp with two crossings so that the resulting knot is alternating. The result is a **twist knot**, pictured in Figure 6.10. Note that the figure-eight knot in Figure 6.5 is an example of a twist knot.

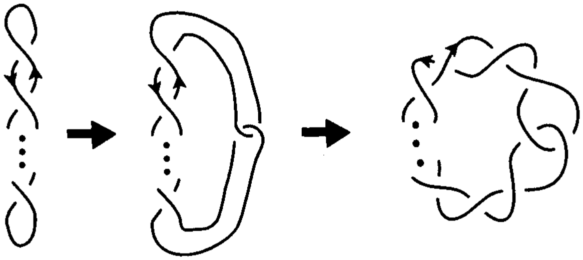


Figure 6.10: A twist knot

Exercise 9. Find a formula for the Conway polynomials of the twist knots. Use this to conclude that the twist knots are distinct for different values of n, and that none of the twist knots are equivalent to $(n, 2)$ torus knots.

6.6 Acknowledgments

The author thanks sarah-marie belcastro and Grant Dasher for suggested revisions and helpful comments.

References

[Ad] Colin C. Adams: *The knot book.* Providence, RI: American Mathematical Society, 2004.

[BQ] Arthur T. Benjamin and Jennifer J. Quinn: *Proofs that really count.* Washington, D.C.: Mathematical Association of America, 2003. (The Dolciani Math. Expositions **27**)

[Cr] Peter R. Cromwell: *Knots and links.* Cambridge: Cambridge Univ. Press, 2004.

[En] Hans Magnus Enzensberger: *The Number Devil.* New York: Henry Holt and Company, 1998.

[Ka] Louis H. Kauffman: The Conway polynomial, *Topology*, **20** #1 (1981), 101–108.

[Ko] Thomas Koshy: *Fibonacci and Lucas numbers with applications.* New York: Wiley-Interscience, 2001. (Pure and Applied Math.)

[Sc] Rob Scharein: Knotplot: Hypnogogic software. `http://knotplot.com`, 1998–2007. (An OpenGL program for visualizing and manipulating mathematical knots.)

FEATURE

7

MATHEMATICAL MINUTIAE

i Has This Funny Property

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

Scott D. Kominers†
Harvard University '09
Cambridge, MA 02138
kominers@fas.harvard.edu

Recall that the **derivative** of a real-valued function $g : \mathbb{R} \to \mathbb{R}$ is given by

$$g'(z) = \lim_{h \to 0} \frac{g(z+h) - g(z)}{h}.$$

While this definition implies continuity, woe be unto the **real analyst**[1] who tries to differentiate the function

$$h_1(x) = \begin{cases} 0 & x \leq 0 \\ x^2 & x \geq 0 \end{cases}$$

twice, even though he may do so once. Indeed, many a differentiable, real-valued function is not twice differentiable.

We define the **complex derivative** of a function $f : \mathbb{C} \to \mathbb{C}$ as

$$f'(z) = \lim_{h \to 0} \frac{f(z+h) - f(z)}{h}$$

where h ranges over complex numbers (if this limit exists). Although this definition looks similar to that of the real derivative, the real and complex derivatives are wildly different beasts. For example, we have the following, which shows that—unlike real-differentiable functions—complex-differentiable functions are always multiply differentiable.

Theorem 1 (Cauchy's Integral Formula [SS, Cor. 4.2] [La, Ch. III Thm. 7.7]). *Any complex function $f : \mathbb{C} \to \mathbb{C}$ that is differentiable near $z_0 \in \mathbb{C}$ is infinitely differentiable there. Furthermore,*

*Zachary Abel, Harvard '10, is a computer science and mathematics concentrator. He is an avid problem solver and researcher, with interests in such varied fields as computational geometry, number theory, partition theory, category theory, and applied origami. He is a founding member of The HCMR and currently serves as Problems Editor, Graphic Artist, and Issue Production Director.

†Scott D. Kominers, Harvard '09, is a mathematics concentrator, ethnomusicology minor, and economics enthusiast. He is an enthusiastic researcher, working in a range of fields including number theory, computational geometry, category theory, mathematical economics, urban economics, law and economics, and historical musicology. He is a founding member of The HCMR and has served as Editor-In-Chief since The HCMR's inception.

[1] real a·na·lyst, ('riːəl 'ænəlɪst), *noun.* (1) A mathematician who studies the analytic properties of the real numbers. (2) An analyst who is not fake.

we may explicitly calculate these values:

$$f^{(n)}(z_0) = \frac{n!}{2\pi i} \oint_\gamma \frac{f(z)}{(z - z_0)^{n+1}}\, dz,$$

where γ is any sufficiently small loop around z_0.

Not only is a complex differentiable function twice differentiable, it is actually **smooth**! Furthermore, it has a convenient power series representation:

Theorem 2 ([SS, Thm. 4.4] [La, Ch. IV Thm. 7.3]). *If f is complex-differentiable near z_0, then f is **analytic** near z_0, i.e. it has a power series expansion*

$$f(z) = \sum_{k=0}^{\infty} c_k (z - z_0)^k$$

in an open neighborhood of z_0.

In light of Theorem 1, we can take Theorem 2 even further, explicitly calculating the power series coefficients: $c_k = \frac{1}{n!} f^{(n)}(z_0) = \frac{1}{2\pi i} \oint_\gamma \frac{f(z)\, dz}{(z - z_0)^{n+1}}$. This is a far cry from the real-analytic case, where even an infinitely differentiable function may not have a power series expansion. The function

$$h_2(x) = \begin{cases} 0 & x \leq 0 \\ e^{-\frac{1}{x}} & x > 0. \end{cases}$$

is a classic example of a smooth function $h_2 : \mathbb{R} \to \mathbb{R}$ with no power series expansion at $x = 0$. (See if you can prove this!)

For any differentiable function $g : \mathbb{R} \to \mathbb{R}$, the image set $g(\mathbb{R})$ is simply an interval.[2] With our trusty i, however, we can see far more about the shape of the image:

Theorem 3 (Liouville's Theorem [SS, Cor. 4.5] [La, Ch. III, Thm. 7.5]). *If $f : \mathbb{C} \to \mathbb{C}$ is an analytic function such that the image $f(\mathbb{C})$ is bounded, then f is constant.*

Whereas in the real case we could only describe the image of a function as an interval, in the complex case we know instead that the image is bounded if and only if it is a single point. We furthermore have control over the images of complex analytic functions with unbounded images:

Theorem 4 (Picard's Little Theorem [SS, Exer. 6.11] [La, Ch. XII, Thm. 2.8]). *If an analytic function $f : \mathbb{C} \to \mathbb{C}$ is nonconstant, then its image omits at most one value. That is, $f(\mathbb{C})$ is either $\mathbb{C}$ or $\mathbb{C} \setminus \{p\}$ for some $p \in \mathbb{C}$.*

So much power in one little i!

Acknowledgements

While there's no i in team, many people contributed to this feature. In particular, i was taught by Andreea Nicoara in the Harvard Mathematics 213a course, Fall 2007, and the authors greatly appreciate Nicoara's instruction.[3] The presentation here also benefited from the comments and advice of Brett Harrison, Daniel Litt, Menyoung Lee, Noam D. Elkies, Shrenik Shah, Ernie Fontes, and Andrea Hawksley. The authors are indebted to the constant advice and support of their parents, {Terry, Bruce} Abel and {Ellen, William} Kominers.

[2] Incidentally, it may be *any* interval—closed, open, or half-open. In fact, this works even if g is only known to be continuous. (Prove it!)

[3] In addition to [La] and [SS], the authors recommend [Al] and [Re], both of which were used as course texts in Nicoara's Mathematics 213a.

References

[Al] Lars V. Ahlfors: *Complex Analysis*, 3rd ed. New York: McGraw-Hill, Inc., 1979.

[La] Serge Lang: *Complex Analysis*, 4th ed. New York: Springer, 1999.

[Re] Reinhold Remmert: *Classical Topics in Complex Function Theory*. New York: Springer, 1998.

[SS] Elias M. Stein and Rami Shakarchi: *Complex Analysis*, Princeton, New Jersey: Princeton University Press, 2003. (Princeton Lectures in Analysis **2**.)

FEATURE

8

STATISTICS CORNER

How Statisticians Discovered the Options Backdating Scandal

Robert W. Sinnott[†]
Harvard University '09
Cambridge, MA 02138
rsinnott@fas.harvard.edu

8.1 Introduction

A **stock option grant** is a contract made between a corporation and another entity in which, on a specified **exercise date**, the corporation agrees to buy or sell a specified number of shares of its stock for a fixed price called the **exercise price**. Such a contract can be quite valuable if the difference between the exercise price and the stock's trading price on the exercise date is large. For example, suppose that a company executive received the option of purchasing 100,000 shares of a company at an exercise price of $35 on January 31. If the stock trades at $55 a share on January 31, then this contract is worth 100,000 × ($55 – $35) = $2,000,000. Thus, it is to the executive's advantage for his options to have a low exercise price, coupled with a high share market price on the exercise date. The latter point is the purpose of the stock option grants—such grants incentivize executives to increase the fundamental value of their companies. The former point is the root of the options backdating debacle.

Options backdating is the practice of marking the grant date of an option with a date prior to date on which the decision to grant the option was made. This is not in general a problem—companies have the right to enter into any agreement and award any compensation according to their internal compensation policies so long as they properly report such awards to their shareholders and the IRS. So that options grants are not counted against company earnings, they must be issued with exercise prices at or above the stock price on the grant date. In the vernacular, these are called **out-of-the-money** (or **at-the-money**) option grants. By contrast, options have a positive value on their grant date are counted against earnings and are called **in-the-money** grants.

Executives desire higher compensation. However, by receiving stock options with grant dates on which stock prices were at a minimum, executives are able to obtain maximal compensation without having to report reduced earnings. Of course, it is difficult to predict stock price minima in the short-term before they occur. Unsurprisingly, however, it is easy to determine such minima *ex post*.

Prior to the 2002 Sarbanes-Oxley reforms, stock option grants could be reported months after the actual grant date, leaving a situation ripe for abuse. The empirical evidence of Lie [L2] demonstrates that, statistically speaking, the probability that such abuse did not occur is impossibly small. Although initial studies such as Yermack [Ye] suggested that executives used insider information to inform the scheduling of option grant dates, Lie's analysis shows conclusively that this explanation is not sufficient to explain the data and that the only reasonable possibility proposed so far is that many option grants were backdated in order to maximize executive compensation. This analysis will be the subject of the remainder of this article.

[†] Robert W. Sinnott, Harvard '09, is a statistics concentrator.

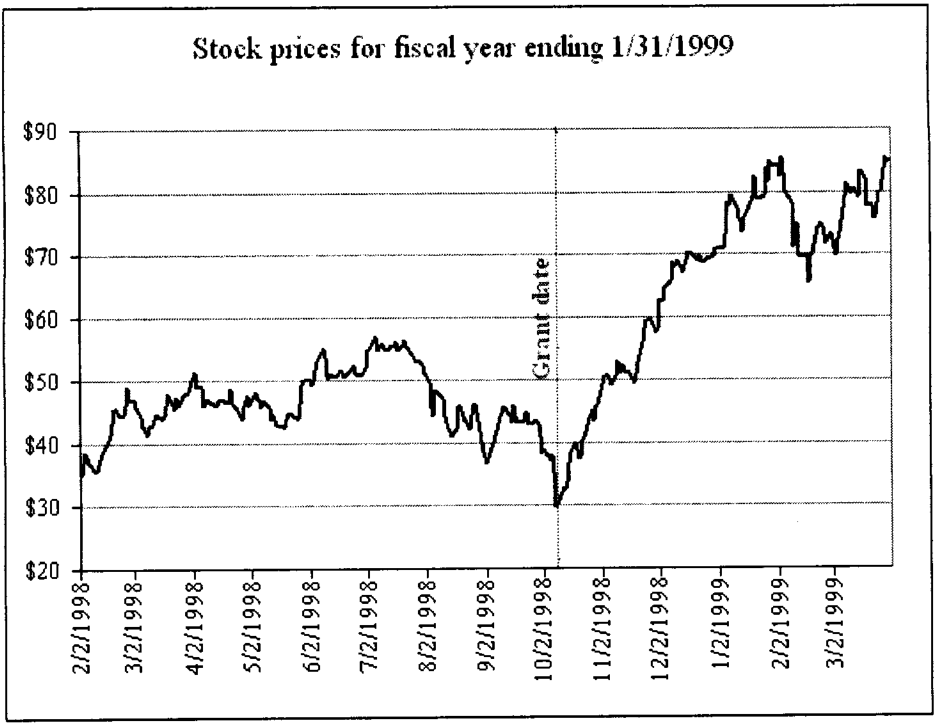


Figure 8.1: Used with permission from Lie [L1].

8.2 Evidence of Option Backdating

Several studies including Chauvin and Shenoy [CS], Yermack [Ye], and Aboody and Kasznik [AK] examined the periods around stock option grants, yielding conflicting results regarding abnormal stock returns before and after stock option grants. However, all of these studies focused on scheduled stock option grants, making their data less reactive to the opportunistic behavior of company executives. Lie [L2] innovatively focused on **unscheduled grants** during the period from 1992–2002, *i.e.*, those grants which were not dated within a week of the grant dates from the previous year.

Lie [L2] sought to answer two questions:

1. Were stock prices on days surrounding stock option grants abnormal?
2. If so, could this abnormality be explained by company executives having inside information about the future returns of their stock?

A negative answer to the second of these questions implies the practice of option backdating.

The first question is reasonably straight forward to answer. Lie used the three factor Fama-French model (see [FF]) to create a baseline for each stock's expected market returns and then compared each company's actual returns to those predicted by the model. This allowed Lie to test whether the stock price was significantly lower at the option grant date than the market would otherwise predict. Abnormal returns in this test would prove that abnormal returns occurred around the grant date but not whether insider information was being used. After all, abnormal returns are being compared to the market, so relative returns are specific to individual companies. The literature has not shown whether or not company executives can accurately predict short term stock price patterns.

The Fama-French three factor model (8.1) was proposed in 1993 by Eugene Fama and Ken French [FF] as a generalization to the widely known Capital Asset Pricing Model (CAPM):

$$R_s - R_f = \beta_1 \times (K_m - R_f) + \beta_2 \times (\text{SMB}) + \beta_3 \times (\text{HML}) + \alpha \tag{8.1}$$

It predicts a stock's returns R_s by a regression of previous returns against overall market returns R_m, the difference between high book-value to price and low book-value to price stock returns HML and the difference in returns between small and large cap stocks SMB (after standardizing using the riskless interest rate R_f). The intercept term for the stock returns is denoted α. As SMB and HML are differenced terms, the R_f standardization cancels out in both cases. This model (8.1) is widely accepted; it is effective as an approximate prediction of returns for individual stocks given knowledge of the returns of other stocks and of general market behavior.

Lie calculated the regression coefficients using the stock price information from the year prior to fifty days before the option grant date. He then used these coefficients to calculate predicted stock prices during the interval surrounding the stock option grant, using standard linear regression techniques. These estimates theoretically allow for a reasonable comparison between the actual stock value and the overall behavior of the market, allowing Lie to isolate firm-specific trends from overall market behavior.

The graph in Figure 8.2 shows the dramatic statistical deviation of average stock prices from their predicted levels around the option grant date for nonscheduled option grants before and after the 2002 Sarbanes-Oxley reforms regulating the use of stock options for executive compensation. The interpretation of the observed patterns is unmistakable. Grants are frequently awarded on dates which correspond to local minima of stock prices and are correlated with the reporting requirements of their corresponding firms. We thus have an answer to the first question: stock prices on days surrounding stock option grants are statistically abnormal.

Having answered the first question, Lie then turned to the second. Could Yermack's [Ye] theory of executive insider information and option grant timing explain the apparent predictive ability of the executives receiving the stock grants? In order to test this, Lie used a logistic regression to determine the factors defining the choice to grant stock options on a particular day, regressing against not only the individual stock's returns but also the returns of the market as a whole. The logic was simple: If executives are working off of inside information, they should be able to predict changes in returns that on the individual firm level but not those on the market level. If, however, the grant appears to be decided not only by the individual stock returns but also by the returns of the market as a whole, the executive options must have been granted *ex post*.

In general, a logistic regression is used when a binary outcome (0 or 1) is being determined. In this case, the binary decision is whether to grant options on a particular date. For the dataset, Lie used the actual option grant dates, and then randomly selected five dates in the six-month range surrounding the option grant for each company (resulting in a total of $10,003$ observations) and set the dependent variable equal to 1 for the actual grant dates and to 0 for the random dates.

For regression coefficients, Lie used both the abnormal (stock-specific) stock market returns and the predicted (market level) stock returns for the intervals of 30–10 days before, 10–5 days before, 5–2 days before, 2–0 days before, 0–2 days after, 2–5 days after, 5–10 days after, and 10–30 days after the grant. He controlled for seasonality by adding dummy variables for each month of the year (eight actual return variables, eight predicted return variables, and eleven month variables in total). In the equation (8.2) below, (MONTHS) is the column vector of seasonal dummy controls, (ABNORMAL) is the column vector of stock-specific returns for each of the listed intervals, and (PREDICT) is the column vector of predicted (market level) stock returns, and $X = (x_{i,j})$ is a data matrix with observations as rows, with the appropriate values for each observation component.

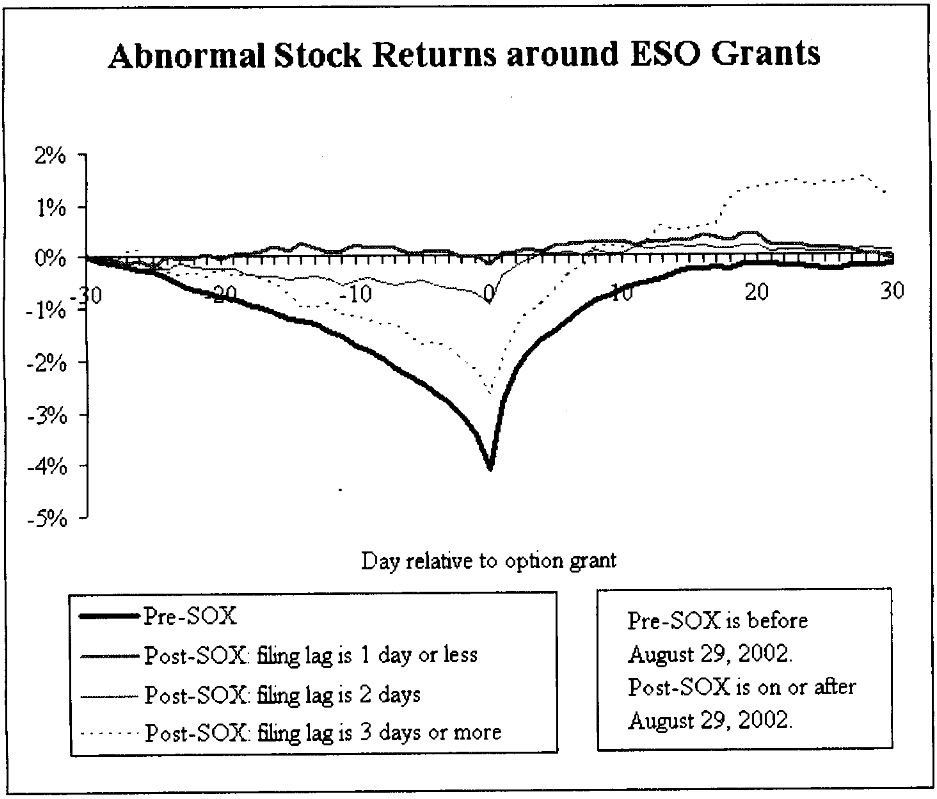


Figure 8.2: Used with permission from Lie [L1].

$$\begin{aligned} \operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + (\text{MONTHS}) \cdot (x_{1,j}, \ldots, x_{11,j}) \\ &+ (\text{ABNORMAL}) \cdot (x_{12,j}, \ldots, x_{19,j}) \\ &+ (\text{PREDICT}) \cdot (x_{20,j}, \ldots, x_{27,j}) \end{aligned} \tag{8.2}$$

Using the estimates derived from this regression, Lie found that the abnormal returns for the regressed intervals to be significant in predicting the "decision" to grant options. He also determined that the overall market predicted returns were very significant in the four days surrounding the option grants. This result conclusively showed that unless executives can effectively predict market level fluctuations in stock price, they must be backdating the options to minimize the grant date price.

8.3 Conclusion

By using standard OLS regression in the Fama-French model and then using logistic regression on the stock returns surrounding the dates of option grants to model the option grant decision, Lie was able to uncover a multi-billion dollar fraud that was occurring in a recently estimated 18.9% of all ESO grants. These simple applications of statistical models have sent shockwaves through the corporate landscape, reinforcing the need for the Sarbanes-Oxley reforms of 2002.

Acknowledgements

The author thanks Erik Lie for his advice and for the use of his graphs.

References

[AK] David Aboody and Ron Kasznick: CEO stock option awards and the timing of corporate voluntary disclosures, *Journal of Accounting Economics* **29** (2001), 73–100.

[CS] Keith W. Chauvin and Catherine Shenoy: Stock price decreases prior to executive stock option grants, *Journal of Corporate Finance* **7** (2001), 53–76.

[FF] Eugene F. Fama and Kenneth R. French: Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* **33** (1993), 3–56.

[L1] Erik Lie: Backdating of executive stock option (ESO) grants, `http://www.biz.uiowa.edu/faculty/elie/backdating.htm`.

[L2] Erik Lie: On the timing of CEO stock option awards, *Management Science* **51** #5 (2005), 802–812.

[Ye] David L. Yermack: Good timing: CEO stock option awards and company news announcements, *Journal of Accounting Economics* **34** (1997), 449–476.

FEATURE

9

APPLIED MATHEMATICS CORNER

Secret Sharing and Applications

Pablo Azar[†]
Harvard University '09
Cambridge, MA 02138
azar@fas.harvard.edu

9.1 Shamir's Secret Sharing Scheme

In this column, I will discuss Shamir's **secret sharing protocol** [Sh]. The motivation for this protocol is the desire for individual privacy while computing an aggregate piece of data. For example, suppose that, to prevent corruption, no employee of a bank is allowed to access the security vault alone. Instead, each employee is given a piece of the password to the vault. When k employees get together, they can reconstruct the password and access the vault. However, $k-1$ of them will have insufficient information about the password, so that no $k-1$ corrupt employees can steal the bank's money.

You can encode a piece of information as a binary ℓ-bit number $a_{\ell-1}2^{\ell-1}+\cdots+2a_1+a_0$ where $a_i \in \{0,1\}$. If the message you want to share is too long, you can split it up into smaller messages so that each takes less than ℓ bits to write. Without loss of generality, we may assume that the message can be written as an ℓ-bit number.

Then we can interpret the message as an element of the finite field $\mathbb{F}_{2^\ell}$.[1] If you have a message $x \in \mathbb{F}_{2^\ell}$, you can compute any polynomial $c_0+c_1x+\cdots+c_nx^n \in \mathbb{F}_{2^\ell}$ and use Lagrange's interpolation theorem: If $\mathbb{F}$ is a field and $(x_1,y_1),\ldots,(x_n,y_n)$ are pairs of points in $\mathbb{F}^2$, then there exists a unique polynomial $p(x)=c_{n-1}x^{n-1}+\cdots+c_0$ of degree $n-1$ such that $p(x_i)=y_i$ for $i=1,\ldots,n$.

What does this have to do with sharing secrets? Well, suppose that you encode a secret s as an ℓ-bit number in $\mathbb{F}_{2^\ell}$ and suppose that you want to distribute s among n people numbered $1,\ldots,n$. You want them to be able to reconstruct s if k people get together and cooperate, but get no information if fewer than k people pool their knowledge. Now, generate $k-1$ random numbers $c_1,\ldots,c_{k-1}$ uniformly from $\mathbb{F}_{2^\ell}$ and consider the polynomial $f(x)=s+c_1x+\cdots+c_{k-1}x^{k-1}$. To each person i, give the **share** $f(i)$.

Players $i_1,\ldots,i_k$ can get together and pool their shares to obtain the set $\{(i_1,f(i_1)),\ldots,(i_k,f(i_k))\}$ of distinct points. With these k points, they can use Lagrange's interpolation theorem to reconstruct the polynomial f. Reconstructing f is equivalent to reconstructing its coefficients. In particular, all k players get knowledge of the secret s, which is the constant coefficient of f.

Furthermore, if $k-1$ players get together they do not learn anything about s. To see this, consider the following argument: given a set of coefficients $(c_1,\ldots,c_{k-1}) \in \mathbb{F}_{2^\ell}^{k-1}$ and a fixed secret s, we can generate the polynomial $f(x)=s+c_1x+\cdots+c_{k-1}x^{k-1}$ and the vector of

[†] Pablo Azar, Harvard 09, is an applied mathematics concentrator from Buenos Aires, Argentina. He is also enrolled in a concurrent masters program in computer science. He is a founding member of The HCMR and currently serves on The HCMRs staff.

[1] You can construct this field if you know an irreducible polynomial $f(x)$ of degree ℓ with coefficients in $\mathbb{F}_2$. The field is given by the quotient $\mathbb{F}_{2^\ell} := \mathbb{F}_2[X]/f(X)$. The elements of this field are polynomials of degree less than ℓ with coefficients in $\mathbb{F}_2$, with all operations conducted modulo $f(X)$. Such polynomials require ℓ bits to encode.

shares $(f(i_1), \ldots, f(i_{k-1})) \in \mathbb{F}_{2^\ell}^{k-1}$. This gives us a map

$$M : \mathbb{F}^{k-1} \to \mathbb{F}^{k-1}$$
$$M(c_1, \ldots, c_{k-1}) = (f(i_1), \ldots, f(i_{k-1})).$$

This map is bijective by Lagrange's interpolation theorem: given a vector of shares $(\alpha_{i_1}, \ldots, \alpha_{i_{k-1}})$, there exists a unique polynomial of degree $k - 1$ with constant coefficient s that interpolates the points $\{(0, s), (i_1, \alpha_{i_1}), \ldots, (i_{k-1}, \alpha_{i_{k-1}})\}$. This polynomial is characterized by its non-constant coefficients $c_1, \ldots, c_{k-1}$.

Therefore, there is a bijection between coefficients $(c_1, \ldots, c_{k-1})$ and shares $(\alpha_1, \ldots, \alpha_{k-1})$. But remember that the dealer chose the coefficients $c_1, \ldots, c_{k-1}$ to be uniformly and independently distributed. This implies that the shares $(\alpha_1, \ldots, \alpha_{k-1})$ are also uniformly and independently distributed. Given the secret s, the players $i_1, \ldots, i_{k-1}$ can get any possible combination of $k - 1$ shares with equal probability. This shows that $k - 1$ shares do not reveal anything valuable about the secret.

9.2 Multi-Party Protocols, Corrupt Players and Corrupt Dealers

The scheme proposed above is very elegant, but the assumptions on the dealer and the honesty of the players may be too strong for applications. The first problem that arises is that there may be no dealer. In this case, each of the players may have a secret $s_1, \ldots, s_n$, and all of them want to compute a function $f(s_1, \ldots, s_n)$ without revealing any information about their corresponding secrets besides what is known from $f(s_1, \ldots, s_n)$ [Ya]. Furthermore, some of the players may be malicious or faulty and give fake or incorrect shares to the other participants. To detect which players are being dishonest, the concept of information checking was introduced by Rabin and Ben-Or [RB]. Their work expands the secret sharing protocol so that, when more than half the players are honest and there are appropriate communication channels, any multiparty computation can be performed by the honest parties.

Another problem may be that of a corrupt dealer. That is, the dealer may be distributing shares $s_1, \ldots, s_n$ to the players so that when players $i_1, \ldots, i_k$ put their shares together, they get the secret s, but when players $j_i, \ldots, j_k$ put their shares together, they get the secret $s' \neq s$. A dealer is honest if and only if the secret reconstructed by any combination of k players is the same. In this case, we say that the players' shares are **consistent**.

To address this second problem, the concept of **Verifiable Secret Sharing** was introduced by Chor, Goldwasser, Micali and Awerbuch [CGMA]. In a Verifiable Secret Sharing scheme, the dealer can broadcast some information, revealing as little information as possible about the shares so that the players can verify that their shares are consistent.· A particularly elegant scheme for doing this was introduced by Feldman [Fe]. In this scheme, the dealer takes a cyclic group G with publicly known generator g, such that obtaining the value of x if one knows g^x is computationally intractable. If $|G|$ is a prime p, all shares in this scheme are in $\mathbb{F}_p$. If the dealer uses the polynomial $f(x) = s + c_1 x + \cdots + c_{k-1} x^{k-1} \bmod p$, she can post $g, g^s, g^{c_1}, \ldots, g^{c_{k-1}}, g^{f(i_1)}, \ldots, g^{f(i_n)}$ on a bulletin board. This way, every player with share $f(i_j)$ can check that $g^{f(i_j)}$ as computed by them is equal to the posted $g^{f(i_j)}$ and all players check that the posted $g^{f(i_j)}$ equals $g^s g^{c_1 i_j} \cdots g^{c_{k-1} i_j^{k-1}}$.[2]

9.3 Two-Party Protocols and Applications

These secret sharing and multi-party computation protocols lead to important applications. A toy example is the **salary problem**, in which n people learn their average salary without revealing anything about their own salaries except what is learned from knowing the sum of all the salaries.

[2] In practice, such groups are constructed by taking primes p, q such that $q = 2p + 1$ and taking $G = \mathrm{Sq}(\mathbb{Z}_q^\times)$, the group of all non-zero squares in the finite field of order q. It is conjectured that there are an infinite number of primes of the form $q = 2p + 1$ where p is prime. Such primes are called **Sophie Germain primes**.

Many applications consider computations with only two parties. Since the multi-party protocol relying on secret sharing needs more than half the players to be honest, it does not apply to two-party computation. Some of these applications (described below), rely on a two-party primitive called **Oblivious Transfer** [Ra, NP], which was introduced by Michael Rabin in 1981.[3] In the Oblivious Transfer protocol, there is one sender and one receiver. The sender has N messages, and the receiver chooses one of them. The protocol is designed so that the sender does not know which message was chosen, and the receiver does not learn anything about any of the other $N - 1$ messages.

One important example is **private querying** of databases. Say Alice has a large database, which Bob pays to use on a per-query basis. However, Bob does not want Alice to know what he is querying. Furthermore, since Alice derives her profit from Bob's queries, she does not want anything revealed to Bob except the results of his query.

Another application is **privacy preserving data mining**. Suppose that two rival companies have datasets D_1 and D_2, on which they want to perform data mining. However, they want to reveal as little as possible about their proprietary data to their rival. Lindell and Pinkas [LP] suggest such a protocol, showing that one can get aggregate data about $D_1 \cup D_2$ revealing as little information as possible about D_1 or D_2 individually. An important lesson from their work is that theoretical protocols may not be the most efficient and that they may need to be modified to accomodate resource constraints. When the databases in question are large, one may want to minimize communication between the parties so as to limit the amount of bandwidth used. The reader interested in the practical applications of multi-party computations is encouraged to look at the report by Du and Atallah [DA], where many interesting problems—including these—are presented.

Shamir's original secret sharing scheme is both simple and applicable. While the generalizations and applications depend on some difficult concepts, the basic secret sharing scheme relies solely on linear algebra in finite fields. It is yet another example demonstrating that mathematics can be modern, elegant, and useful.

Acknowledgements

The vast majority of the information presented here was demonstrated to me in some form by Professor Michael O. Rabin of Harvard University. I would also like to thank Shrenik Shah, for his extensive edits. The exposition of Shamir's secret sharing scheme owes much to his guidance.

References

[CGMA] Benny Chor, Shafi Goldwasser, Silvio Micali, Baruch Awerbuch: Verifiable Secret Sharing in the Presence of Faults, *Proc. 26th IEEE Symp. on Foundations of Computer Science* (1985).

[DA] Wenliang Du, Mikhail J. Atallah: Secure Multi-Problem Computation Problems and Their Applications: A Review and Open Problems, *CERIAS Tech Report 2001-51* (2001).

[Fe] Paul Feldman: A practical scheme for non-interactive verifiable secret sharing. *Proceedings of the 28th IEEE Symposium on the Foundations of Computer Science* (1987).

[LP] Yehuda Lindell, Benny Pinkas: Privacy Preserving Data Mining. *J. Cryptology* **15** (2002), 177–206.

[NP] Moni Naor, Benny Pinkas: Oblivious transfer and polynomial evaluation, *Proceedings of the thirty-first annual ACM symposium on Theory of computing* (1999).

[Ra] Michael O. Rabin: How to exchange Secrets with Oblivious Transfer, *Technical Report TR-81* Harvard University: Aiken Computation Lab (1981).

[3] The version I am presenting is given by Naor and Pinkas [NP]

[RB] Tal Rabin, Michael Ben-Or: Verifiable secret sharing and multiparty protocols with honest majority, *Proceedings of the twenty-first annual ACM symposium on Theory of computing* (1989).

[Sh] Adi Shamir: How to share a secret, *Communications of the ACM* (1979).

[Ya] Andrew C. Yao: Protocols for secure computations, *Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science* (1982), 160–164.

FEATURE

10

MY FAVORITE PROBLEM

Linear Independence of Radicals

Iurie Boreico[†]
Harvard University '11
Cambridge, MA 02138
boreico@fas.harvard.edu

The problem I intend to discuss here was mentioned in the prior *HCMR*—in particular, author Zachary Abel [Ab, p. 79] stated that "the set $\{\sqrt{n} \mid n \in \mathbb{N}$ is squarefree$\}$ is linearly independent over rationals." More formally:

Problem. *Let $n_1, n_2, \ldots, n_k$ be distinct squarefree integers. Show that if $a_1, a_2, \ldots, a_k \in \mathbb{Z}$ are not all zero, then the sum $S = a_1\sqrt{n_1} + a_2\sqrt{n_2} + \ldots + a_k\sqrt{n_k}$ is non-zero.*

Note that this problem is equivalent to Abel's statement, since we may clear denominators to obtain coefficients in $\mathbb{Z}$.

10.1 Preliminary Analysis

By setting $A_i = a_i^2 n_i$, the problem can be restated as follows: if

$$\sum_{i=1}^{k} \pm\sqrt{A_i} = 0,$$

then at least one of the expressions A_i/A_j must be a perfect square. Indeed, in our case none of the expressions $A_i/A_j = (a_i/a_j)^2 n_i/n_j$ is a perfect square, so the sum

$$a_1\sqrt{n_1} + a_2\sqrt{n_2} + \ldots + a_k\sqrt{n_k} = \sum_{i=1}^{k} \pm\sqrt{A_i}$$

must not be zero. The converse follows similarly.

In this form, the problem can be tackled for small values of k by simply squaring. For example, if $k = 2$, we have $\sqrt{A_1} - \sqrt{A_2} = 0$, so $\sqrt{A_1} = \sqrt{A_2}$. Squaring gives $A_1 = A_2$ and thus $A_1/A_2 = 1$, which is a perfect square. If $k = 3$ we have WLOG that $\sqrt{A_1} = \pm\sqrt{A_2} + \sqrt{A_3}$, and so again squaring gives $A_1 = A_2 + A_3 \pm 2\sqrt{A_2A_3}$. Hence $\sqrt{A_2A_3} = \pm(A_2 + A_3 - A_1)/2$, which implies $A_2A_3 = (A_2 + A_3 - A_1)^2/4$ is a perfect square, and so

$$\frac{A_2}{A_3} = \frac{A_2A_3}{A_3^2} = \left(\frac{A_2 + A_3 - A_1}{2A_3}\right)^2.$$

For $k = 4$ we may rewrite the problem as $\pm\sqrt{A_1} \pm \sqrt{A_2} = \pm\sqrt{A_3} \pm \sqrt{A_4}$. Then by squaring we have $A_1 + A_2 - A_3 - A_4 \pm 2\sqrt{A_1A_2} \pm 2\sqrt{A_3A_4} = 0$ and we handle this using the previously established case $k = 3$.

[†] Iurie Boreico, Harvard '11, is a prospective mathematics concentrator residing in Weld. His mathematical knowledge is yet too vague to define his interests, but they tend towards number theory. When not doing math, he usually misses his home country, Moldova, or wastes his time in some other way.

Unfortunately, this approach does not extend to $k > 4$, for however we rearrange the given expressions, squaring only *increases* the number of radicals. In fact, as an olympiad-style problem, this problem is very hard, and probably very few would be successful in solving it. With enough patience and creativity, however, several solutions are possible.

10.2 Solutions

Solution 1 from [Kv]. Let $p_1, p_2, \ldots, p_N$ be all the primes dividing $n_1 n_2 \cdots n_k$. We will prove the following statement by induction on N:

Recall that $S = a_1\sqrt{n_1} + a_2\sqrt{n_2} + \ldots + a_k\sqrt{n_k}$. Then there exists an expression $S' = b_1\sqrt{m_1} + b_2\sqrt{m_2} + \ldots + b_l\sqrt{m_l}$ where $m_1, m_2, \ldots, m_l$ are squarefree integers with prime factors among the $p_1, p_2, \ldots, p_N$ and b_i are integers, such that SS' is a non-zero integer (in particular, $S \neq 0$, as desired).

For $N = 0$ this is obvious, as in this case $k = 1, n_1 = 1$ and we get $S = a_1 \neq 0$ so we can set $S' = 1$. For $N = 1$ we either have $S = a_1\sqrt{p_1}$, in which case we may let $S' = \sqrt{p_1}$, or we have $S = a_1\sqrt{p_1} + a_2$. In the last case we may take $S' = -a_1\sqrt{p_1} + a_2$, so $SS' = a_2^2 - a_1^2 p_1$. This is non-zero as a_2^2 is divisible by an even power of p_1, whereas $a_1^2 p_1$ is divisible by an odd power of p_1, so the two cannot be equal.

Now we perform the induction step. Assume that the theorem is true for $N \leq n$; we prove it for $N = n + 1$. Let $p_N = p_{n+1} = p$. We may write $S = S_1 + S_2\sqrt{p}$ where the primes appearing in the radicals in S_1, S_2 are among $p_1, \ldots, p_n$, and further, neither S_1 nor S_2 is identically 0 (else we would already be done, as p would be irrelevant). So there exists a sum S_2' of the form given in the claim such that $S_2 S_2'$ is a non-zero integer k.

The intermediate product SS_2' can then be written as $S_4 + k\sqrt{p}$ where the primes appearing in the radicals in S_4 are also among $p_1, \ldots, p_n$. We may thus multiply it by $S_4 - k\sqrt{p}$ to get $S_4^2 - k^2 p$. Finally, it is easy to see that all prime factors of radicals of $S_4^2 - k^2 p$ are among $p_1, \ldots, p_n$, so, assuming this number is not itself zero, the induction hypothesis implies that there exists a non-zero weighted sum of radicals S_5 whose prime factors appear among $p_1, p_2, \ldots, p_n$ such that $(S_4^2 - k^2 p)S_5$ is a non-zero integer. So we obtain the desired representation $SS_2'(S_4 - k\sqrt{p})S_5 \in \mathbb{Z} \setminus \{0\}$ where $S' = S_2'(S_4 - k\sqrt{p})S_5$ is a sum of radicals of the desired type.

Thus, we are done if we manage to prove we do not run into trouble when multiplying $S_4 - k\sqrt{p}$, as the product could become zero at that step (e.g. if $S_4 - k\sqrt{p} = 0$). It is sufficient to prove that $S_4^2 - k^2 p \neq 0$. If S_4 is an integer this is clear, and if S_4 is of form $u\sqrt{q}$ this also true because $u^2 q \neq k^2 p$, as p does not divide q. Otherwise, S_4 contains at least two distinct radicals in its canonical expression (if we consider $\sqrt{1}$ as a radical),[1] and we can assume without loss of generality that p_n appears in one of these two radicals but not in the other. So $S_4 = S_6 + S_7\sqrt{p_n}$ where S_6, S_7 are sums of radicals with prime factors among $p_1, p_2, \ldots, p_{n-1}$, and $S_6, S_7 \neq 0$. Therefore $S_4^2 - k^2 p = S_6^2 + 2S_6 S_7\sqrt{p_n} + S_7^2 p_n - k^2 p$. As $S_6 S_7 \neq 0$, by expanding the expression $S_4^2 - k^2 p$ we will get at least one radical containing p_n. But then by the induction hypothesis, the expression is non-zero as claimed. □

This solution, even if it might seem somewhat unnatural and tedious, is completely logical in its construction. By starting from the well-known idea of multiplication by a conjugate (the case $N = 1$ above), the idea is to actually produce a sort of "conjugate" expression for more complicated sums involving radicals, *i.e.* something involving the same radicals which when multiplied by the original produces an integer. The (somewhat unappealing) induction step is just a set of technical manipulations that help realize this idea.

[1] By the canonical representation of an expression involving radicals we mean its simplest possible form—that is, the form obtained by extracting the squares out of the radicals and grouping together the terms which have the same square-free numbers under the radicals. For example, $(\sqrt{2} + \sqrt{10})^2 = 2 + 2\sqrt{2} \cdot \sqrt{10} + 10$ would be brought to $2 + 4\sqrt{5} + 10 = 12 + 4\sqrt{5}$. The statement of the problem is just the fact that the canonical representation is indeed "canonical," that is, the same number can not be written as such a sum in two different ways (otherwise subtracting the two expressions would produce a counterexample).

If one takes as a starting point the mentioned idea of conjugate expressions, one might consider the following question:

Question. *What is the expression conjugate to* $\sqrt{a_1} + \sqrt{a_2} + \ldots + \sqrt{a_n}$*?*

We know that for $n = 2$ the conjugate is $\sqrt{a_1} - \sqrt{a_2}$. Of course we could have chosen some other combination of signs, like $-\sqrt{a_1} - \sqrt{a_2}$ or $-\sqrt{a_1} + \sqrt{a_2}$, but we do not get anything new from them, as these two expression are just the original ones with the opposite sign. Given this example, we might thnk that the expression $\sqrt{a_1} + \sqrt{a_2} + \ldots + \sqrt{a_n}$ has many conjugates, and that they represent all expressions of form $\pm\sqrt{a_1} \pm \sqrt{a_2} \pm \ldots \pm \sqrt{a_n}$ for all combinations of pluses and minuses. Again, we need to ensure that the same expression does not occur twice, the second time with opposite sign, which can be realized by requiring that the sign of $\sqrt{a_1}$ is always positive. We get a family of 2^{n-1} alike sums: $\sqrt{a_1} \pm \sqrt{a_2} \pm \sqrt{a_3} \pm \ldots \pm \sqrt{a_n}$. This might suggest that the product of this entire family could in fact be the required non-zero integer we sought in Solution 1, but unfortunately while it is indeed possible that this product is an integer, there is no obvious way to handle this huge expression directly and prove that it is non-zero.

These considerations inspire the next solution:

Solution 2. Consider the linear expression $L(x_1, x_2, \ldots, x_n) = a_1x_1 + a_2x_2 + \ldots + a_nx_n$. We will also consider its conjugate expressions of form $L'(x_1, x_2, \ldots, x_n) = a_1x_1 \pm a_2x_2 \pm a_3x_3 \pm \ldots \pm a_nx_n$. There are 2^{n-1} such expressions. Now take a variable T and consider the polynomial

$$F_{L,x_1,x_2,\ldots,x_n}(T) = \prod_{L'} (T - L'(x_1, x_2, \ldots, x_n)) = \prod (T - a_1x_1 \pm a_2x_2 \pm \ldots \pm a_nx_n),$$

where the product is taken over all conjugate expressions L' (including L).

Note that $F_{L,x_1,x_2,\ldots,x_n}(T)$ is written as a polynomial in T, but can be considered as a polynomial in $x_1, x_2, \ldots, x_n$. Also note that changing the signs of any of $x_2, x_3, \ldots, x_n$ will not affect F because doing so only permutes the set $\{L'\}$. Therefore

$$F_{L,x_1,x_2,\ldots,x_n}(T) = F_{L,x_1,\pm x_2,\pm x_3,\ldots,\pm x_n}(T).$$

In particular, if we expand the product representation of F into a sum of monomials, each monomial term will contain only even powers of x_k ($k = 2, \ldots, n$), because otherwise changing the sign of x_k would change the sign of the monomial. Note that this is not true for x_1, as doing so sends the set $\{L'\}$ to the set $\{-L'\}$. But by expanding F into a sum of monomials and grouping the monomials with odd and even powers of x_1 we can write

$$F_{L,x_1,x_2,\ldots,x_n}(T) = x_1P(x_1^2, x_2, x_3, \ldots, x_n, T) + Q(x_1^2, x_2, x_3, \ldots, x_n, T).$$

As we have seen above, P and Q do involve only monomials with even powers of $x_2, x_3, \ldots, x_n$, and so they depend only on $x_2^2, x_3^2, \ldots, x_n^2$. So we can actually write

$$F_{L,x_1,x_2,\ldots,x_n}(T) = x_1P_2(x_1^2, x_2^2, x_3^2, \ldots, x_n^2, T) + Q_2(x_1^2, x_2^2, x_3^2, \ldots, x_n^2, T).$$

It is also clear that if a_i are integers then all the coefficients of P and Q will be integers.

Now let us return to the problem. We actually prove a different version of it: that is, that no non-zero integer M can be represented as a nontrivial canonical sum of radicals. To see that this implies the original problem, assume that $\sum_{i=1}^{k} a_i\sqrt{n_i} = 0$. Then, by multiplying by $\sqrt{n_k}$, we get $\sum_{i=1}^{k-1} a_i\sqrt{n_in_k} = -a_kn_k$, which is a contradiction if we prove that no non-zero integer can be represented as a canonical sum of radicals. So let us prove this version, by induction on k. The base case, $k = 1$, is clear.

If we assume that an expression of form $a_1\sqrt{n_1} + a_2\sqrt{n_2} + \ldots + a_k\sqrt{n_k}$ equals $M \in \mathbb{Z}\setminus\{0\}$, then the polynomial $F_{L,\sqrt{n_1},\sqrt{n_2},\ldots,\sqrt{n_k}}(T)$ would vanish at $T = M$. But we saw the polynomial can be written simply as

$$\sqrt{n_1}P_2(n_1, n_2, \ldots, n_k, T) + Q_2(n_1, n_2, \ldots, n_k, T),$$

so we would have $\sqrt{n_1}P_2(n_1,n_2,\ldots,n_k,M)+Q_2(n_1,n_2,\ldots,n_k,M)=0$. But $P_2(n_1,\ldots,n_k,T)$ and $Q_2(n_1,\ldots,n_k,T)$ are integers. By the base case, $A+B\sqrt{n_1}=0$ implies $A=B=0$, so we have

$$P(n_1,n_2,\ldots,n_k,M)=Q(n_1,n_2,\ldots,n_k,M)=0.$$

Hence,

$$-\sqrt{n_1}P(n_1,n_2,\ldots,n_k,M)+Q(n_1,n_2,\ldots,n_k,M)=0,$$

i.e. $F_{M,-\sqrt{n_1},\sqrt{n_2},\ldots,\sqrt{n_k}}(M)=0$. Thus,

$$\prod(M+a_1\sqrt{n_1}\pm a_2\sqrt{n_2}\pm a_3\sqrt{n_3}\pm\ldots\pm a_k\sqrt{n_k})=0,$$

and so $M=-a_1\sqrt{n_1}\pm a_2\sqrt{n_2}\pm\ldots\pm a_k\sqrt{n_k}$ for some combination of signs. However, we already have $M=a_1\sqrt{n_1}+a_2\sqrt{n_2}+\ldots+a_k\sqrt{n_k}$, and summing these two equalities gives

$$2M=(a_2\pm a_2)\sqrt{n_2}+(a_3\pm a_3)\sqrt{n_3}+\ldots+(a_k\pm a_k)\sqrt{n_k}.$$

This cannot happen by the induction hypothesis, and we have reached our contradiction. □

10.3 Further Ideas

Now I will explain why I like this problem. The essential reason is that the solutions hint at many important concepts in algebra and number theory. let us talk about some of them:

The primitive element theorem. The primitive element theorem states that any finite separable field extension L/K contains a primitive element, *i.e.* an element that generates the whole extension. This problem allows us to explicitly find a primitive element (in fact many of them) for the extension $\mathbb{Q}[\sqrt{p_1},\sqrt{p_2},\ldots,\sqrt{p_k}]/\mathbb{Q}$, where $p_1,p_2,\ldots,p_k$ are distinct primes.

The field $\mathbb{Q}[\sqrt{p_1},\sqrt{p_2},\ldots,\sqrt{p_k}]$ consists of all combinations $\sum a_i\sqrt{n_i}$ where $a_i\in\mathbb{Q}$ and $n_1,n_2,\ldots,n_{2^k}$ are all possible products that can be formed with $p_1,p_2,\ldots,p_k$. As we have proven that $\sqrt{n_1},\sqrt{n_2},\ldots,\sqrt{n_k}$ are independent over $\mathbb{Q}$, it follows that the degree of the extension over $\mathbb{Q}$ is 2^k. Thus to find a primitive element means to find an element θ in $\mathbb{Q}[\sqrt{p_1},\ldots,\sqrt{p_k}]$ which is a root of an irreducible polynomial of degree 2^k. We claim our friend $\sqrt{p_1}+\sqrt{p_2}+\ldots+\sqrt{p_k}$ (or any of its conjugates) can be taken as θ.

Assume $P\in\mathbb{Q}[X]$ and $P(\theta)=0$, with P irreducible over $\mathbb{Q}$. We can expand each power of θ in $P(\theta)$ and write it as a sum of radicals, and then combine these radicals to obtain $P(\theta)=a_1\sqrt{n_1}+a_2\sqrt{n_2}+\ldots+a_{2^k}\sqrt{n_k}$. The results proven above tell us that $P(\theta)=0$ if and only if all a_i are 0. Let us take now a conjugate of θ, say $\theta'=\epsilon_1\sqrt{p_1}+\epsilon_2\sqrt{p_2}+\ldots+\epsilon_k\sqrt{p_k}$ where $\epsilon_i\in\{-1,1\}$. We claim $P(\theta')=0$. Indeed, if we expand θ'^k, the coefficient of $\sqrt{n_i}$ will either be the same as the coefficient of $\sqrt{n_i}$ in θ^k, or it will be the additive inverse of that coefficient, depending on how many of p_j with $\epsilon_j=-1$ divide $\sqrt{n_i}$. We thus get $P(\theta')=\sum_i a_i\mu_i\sqrt{n_i}$ where $\mu_i=\prod_{p_j|n_i}\epsilon_j$. As all a_i are zero, we get $P(\theta')=0$. Hence P has as roots all the conjugates of $\pm\theta$, of which there are 2^k, so P has degree at least 2^k. In fact it must have degree 2^k because θ lies in an extension of degree 2^k, so θ is indeed a primitive element.

It is also clear now that $P(X)=\prod_{\epsilon_i\in\{-1,1\}}(X-\epsilon_1\sqrt{p_1}-\epsilon_2\sqrt{p_2}-\ldots-\epsilon_k\sqrt{p_k})$. The fact that this polynomial has rational coefficients follows from the fact that

$$P(X)=F_{L,\sqrt{p_1},\sqrt{p_2},\ldots,\sqrt{p_k}}(X)\cdot F_{L,-\sqrt{p_1},\sqrt{p_2},\ldots,\sqrt{p_k}}(X)$$

(we keep the notations of Solution 2) and this has integer coefficients from Solution 2.

The degree of extensions of radicals. We noted above that $[\mathbb{Q}(\sqrt{p_1},\sqrt{p_2},\cdots,\sqrt{p_k}):\mathbb{Q}]=2^k$ when $p_1,p_2,\ldots,p_k$ are distinct primes. It would be interesting to show that $[\mathbb{Q}(\sqrt{p_1},\sqrt{p_2},\ldots,\sqrt{p_k}):\mathbb{Q}]=2^k$ with a constraint weaker than that p_i be primes. The degree of this extension is trivially at most 2^k, but it may be less than that. For example we might have $\sqrt{p_3}\in\mathbb{Q}(\sqrt{p_1},\sqrt{p_2})$ if $\sqrt{p_3}=m\sqrt{p_1p_2}$, where $m\in\mathbb{Q}$, in which case adjoining p_3 would not alter the extension. We will prove that these are exactly the "uncomfortable" cases. Namely, let us take from $p_1, p_2, \ldots, p_k$ a maximal sequence $p_1, \ldots, p_l$ which is multiplicatively independent, by which we mean that the product of any nonempty subset of elements in the sequence is not a perfect square. More explicitly, $p_1, p_2, \ldots, p_l$ is multiplicatively independent, but for any $j > l$, there exist $1 \le i_1, \ldots, i_r \le l$ such that $p_j p_{i_1} p_{i_2} \cdots p_{i_r} = a^2$ is a perfect square. This means

$$\sqrt{p_j} = \frac{a}{p_{i_1} p_{i_2} \cdots p_{i_r}} \sqrt{p_{i_1} p_{i_2} \cdots p_{i_r}} \in \mathbb{Q}(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_l}),$$

and so $\mathbb{Q}(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_l}, \sqrt{p_j}) = \mathbb{Q}(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_l})$. It is easy to see that $[\mathbb{Q}(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_l}) : \mathbb{Q}] = 2^l$, by arguments similar to those in previous paragraph, as the numbers $p_1, p_2 \ldots p_l$ have distinct squarefree parts and so are linearly independent over $\mathbb{Q}$. (If they did not, they could be multiplied to obtain perfect squares). Also, as above, $\sqrt{p_1} + \sqrt{p_2} + \ldots + \sqrt{p_l}$ is a primitive element of the extension, so the primitive element theorem is verified explicitly in this case.

Note that $\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_k}$ with the operation of multiplication (and division) generate an abelian group A. Let $\mathbb{Q}^\times$ be the multiplicative group $(\mathbb{Q}, \cdot)$. The group $G = A\mathbb{Q}^\times$ satisfies $[G : \mathbb{Q}^\times] = 2^l$, with $\sqrt{n_1}, \sqrt{n_2}, \ldots, \sqrt{n_l}$ forming a complete set of representatives for $G/\mathbb{Q}^\times$, where the n_i are all the possible 2^l products $\prod_{i \in J \subset \{1,2,\ldots,l\}} p_i$. (The quotient $G/\mathbb{Q}^\times$ is generated by the images of $\sqrt{n_1}, \sqrt{n_2}, \ldots, \sqrt{n_l}$ and is isomorphic to $(\mathbb{Z}/2\mathbb{Z})^l$). Therefore the result obtained in this section can be rewritten as $[\mathbb{Q}(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_k}) : \mathbb{Q}] = [G : \mathbb{Q}^\times]$. In this abstract form, the result is easier to generalize.

Galois groups. The freedom with which one interchanged signs in front of radicals may suggest in fact that there is no visible difference between $\sqrt{p}$ and $-\sqrt{p}$, and they can be interchanged in expressions when one is concerned with rationals. This idea leads to the Galois groups. Indeed, if $p_1, p_2, \ldots, p_l$ are multiplicatively independent then in any expression $F(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_k})$ with $F \in \mathbb{Q}[X_1, X_2, \ldots, X_k]$ one may change the signs to get $F(\pm\sqrt{p_1}, \pm\sqrt{p_2}, \ldots, \pm\sqrt{p_k})$, and the new expression will be conjugate to the original. In particular, it will equal 0 if and only if the original equals 0. We thus have 2^l isomorphisms of $\mathbb{Q}(\sqrt{p_1}, \sqrt{p_2}, \sqrt{p_3}, \ldots, \sqrt{p_l})$ for any choices of signs $\epsilon_1, \epsilon_2, \ldots, \epsilon_l \in \{-1, 1\}$, characterized by sending $\sqrt{p_i}$ to $\epsilon_i \sqrt{p_i}$. As the extension is normal (since $\mathbb{Q}(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_l})$ is the splitting field of $(X^2 - p_1)(X^2 - p_2) \ldots (X^2 - p_l))$ and has degree 2^l, we have found the Galois group of the extension: $(\mathbb{Z}/2\mathbb{Z})^l$.

Higher powers. The natural question is whether the statement of the problem can be extended to radicals of any degree. Specifically, we prove that if $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n \in \mathbb{Q}^+$ and $\sqrt[k_i]{b_i}$ are not all rational, then $\sum_{i=1}^n a_i \sqrt[k_i]{b_i}$ is not rational. The solution is a generalization of Solution 2. Firstly, we may assume all the k_i equal, as otherwise we can replace them by their least common multiple and adjust the b_i accordingly. So we need to prove that a sum of form $\sum_{i=1}^n a_i \sqrt[k]{b_i} = M \in \mathbb{Z}$ cannot occur if at least one of the b_i is not a perfect k-th power. Again, we use induction on n.

Consider ξ a primitive k-th root of unity. Take the polynomial

$$P(X, b) = \prod \left(X - b - \xi^{i_2} a_2 \sqrt[k]{b_2} - \ldots - \xi^{i_n} a_n \sqrt[k]{b_n} \right)$$

with a_i rational, where the product is taken over all choices of $i_2, i_3, \ldots, i_n \in \{0, 1, \ldots, k-1\}$. As replacing $\sqrt[k]{b_i}$ by $\xi \sqrt[k]{b_i}$ in the above expression preserves P, we conclude that P can be written as a polynomial in X and b with coefficients in $\mathbb{Q}[b_2, b_3, \ldots, b_n] = \mathbb{Q}$ (we did not argue this rigorously since it is completely similar to the argument used in the original problem). Now if $M \in \mathbb{Q}$ can be written as $\sum_{i=1}^n a_i \sqrt[k]{b_i}$ then $P(M, a_1 \sqrt[k]{b_1}) = 0$. Let $d \mid k, d > 1$ be the smallest integer such that $\sqrt[k]{b_1^d} \in \mathbb{Q}$; then $P(M, x)$ can be written as

$$q_0(x^d) + x q_1(x^d) + \ldots + x^{d-1} q_{d-1}(x^d)$$

where $q_0, q_1, \ldots, q_{d-1} \in \mathbb{Q}[X]$. So if $(a_1 \sqrt[k]{b_1})^d = u \in \mathbb{Q}$ we have

$$q_0(u) + q_1(u) \sqrt[d]{u} + \ldots + q_{d-1}(u) \sqrt[d]{u^{d-1}} = 0.$$

Now note that $1, \sqrt[d]{u}, \ldots, \sqrt[d]{u^{d-1}}$ are independent over $\mathbb{Q}$, because $\sqrt[d]{u}$ is the root of the irreducible polynomial $X^d - u$. (To see that this polynomial is irreducible, note that the roots of $X^d - u$ have absolute value $\sqrt[d]{|u|}$, so if $f(x) \mid X^d - u$ has degree m, then $|f(0)| = \sqrt[d]{|u|^m}$. But for $0 < m < d$ this is not rational, for if it were then $(a_1 \sqrt[k]{b_1})^m = \pm \sqrt[d]{|u|^m}$ would be rational, contradicting the minimality of d. So $X^d - u$ does not have proper factors in $\mathbb{Q}[X]$ and is irreducible.) Therefore we conclude that $q_0(u) = q_1(u) = q_2(u) = \ldots = q_{d-1}(u) = 0$. If ϵ is a primitive d-th root of unity we may conclude that

$$P(M, \epsilon u) = q_0(u) + q_1(u)\epsilon \sqrt[d]{u} + \ldots + q_{d-1}(u)\epsilon^{d-1} \sqrt[d]{u^{d-1}} = 0,$$

so $M = \epsilon \sqrt[d]{u} + \sum_{i=2}^{n} \xi^{l_i} a_i \sqrt[k]{b_i}$ for some $\{l_i\}$. But then

$$\epsilon \sqrt[d]{u} + \sum_{i=2}^{n} \xi^{l_i} a_i \sqrt[k]{b_i} = \sqrt[d]{u} + \sum_{i=2}^{n} a_i \sqrt[k]{b_i}.$$

This is impossible as each of the terms in the left-hand side has real part less than or equal to the corresponding term of the right-hand side (which a positive real of the same absolute value), and the inequality is strict for the first term.

References

[Ab] Zachary Abel: My Favorite Problem: Bert and Ernie, *The Harvard College Mathematics Review* **1** #2 (2007), 78–83.

[Kv] Irrationality of a Sum of Radicals (in Russian), *Kvant* **2** (1972).

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 *September 15, 2008*. An asterisk beside a problem or part of a problem indicates that no solution is currently available.

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).

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 rooms, each of which is empty except for $s > 0$ switches. 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 (Walt Whitman HS '08), and Justin Chen (Caltech '09).

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).

[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.

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).

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).

The following problem from the Fall 2007 issue received no submissions. Since this problem defied solution, we are rereleasing it for one more issue.

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.

FEATURE 12
Solutions

Projective Paranoia

S07 – 3. The incircle Ω_{ABC} of a triangle ABC is tangent to BC, CA, AB at P, Q, R respectively. Rays PQ and BA intersect at M, rays PR and CA intersect at N, and the incircle Ω_{MNP} of triangle MNP is tangent to MN and NP at X and Y respectively. Given that X, Y and B are collinear, prove:

(a) Circles Ω_{ABC} and Ω_{MNP} are congruent, and

(b) these circles intersect each other in 60° arcs.

Proposed by Zachary Abel '10.

Solution by the proposer. Let $\text{cr}(A_1, A_2; A_3, A_4)$ denote the cross ratio $\frac{A_1A_3/A_3A_2}{A_1A_4/A_4A_2}$ of four collinear points A_1, A_2, A_3, A_4. Let ρ and τ denote the polar maps through circles Ω_{ABC} and Ω_{MNP} respectively, and let I and J be the respective centers of these two circles.

Let Ω_{MNP} touch MP at Z, and define $MN \cap BC = S$. We first show that X, Z, and C are collinear. As $\rho(S) = AP$, it follows that $\text{cr}(R, Q; S, RQ \cap AB) = -1$, and hence by perspectivity through A, $\text{cr}(B, C; S, P) = -1$. An identical argument proves that $\text{cr}(N, M, MN \cap YZ, X) = -1$. Letting $C' = XZ \cap BC$, we may calculate

$$\text{cr}(B, C'; S, P) \overset{X}{=} \text{cr}(Y, Z; MN \cap YZ, XP \cap YZ) \overset{P}{=} \text{cr}(N, M; MN \cap YZ, X) = -1$$

where the notation $\overset{J}{=}$ indicates that equality follows by a perspectivity about point J. Since $\text{cr}(B, C; S, P) = \text{cr}(B, C'; S, P)$, it follows that $C = C'$, as claimed.

By Pascal's theorem on hexagon $BXCNPM$, we find that Y, A, and Z are collinear. Then by the converse of Brianchon's theorem on hexagon $RYNMZQ$, there must be some conic tangent to line RN at Y, tangent to QM at Z, and tangent to lines MN and QR. Such a conic is uniquely determined by the first three of these four facts and must therefore coincide with circle Ω_{MNP}, hence QR is tangent to this circle. Label this point of tangency T. By well-known properties of circumscribed quadrilaterals, lines XY, MR, and ZT are concurrent, *i.e.* T lies on line ZB. Likewise, T is collinear with Y and C.

We may apply similar arguments to circle Ω_{ABC}. Letting $U = CT \cap AB$ and $V = BT \cap AC$, the converse of Brianchon's Theorem proves that UV is tangent to circle Ω_{ABC} at some point W. Also as above, RW must pass through Z and QW passes through Y.

Note that

$$\rho(UC) = \rho(U) \cap \rho(C) = RW \cap PQ = Z$$

and likewise $\rho(VB) = Y$, so $\rho(T) = \rho(UC \cap VB) = YZ$. In particular, $IT \perp YZ$, and hence $IT \parallel PJ$. Similarly, $\tau(A)$ is the join of $\tau(NQ) = XY \cap TZ = B$ and $\tau(MR) = XZ \cap YT = C$, hence $JA \perp BC$. Let line JA meet BC and the top of circle Ω_{MNP} at D and E respectively.

By Pascal's theorem on hexagon $PPQQRR$ in circle Ω_{PQR}, point S lies on line QR. As $\angle JTS = \angle JDS = 90°$, quadrilateral $JTDS$ is cyclic. Using α, β, and γ for the angles of $\triangle ABC$, we may calculate

$$\angle JET = \tfrac{1}{2}\angle DJT = \tfrac{1}{2}\angle DST = \tfrac{1}{2}\,(180° - \angle RQP - \angle QPS) = 45° - \tfrac{\alpha}{4} - \tfrac{\gamma}{2}$$

and, with $YZ \cap BC = F$,

$$\angle PFZ = \angle CPZ - \angle FZP = (90° - \tfrac{\gamma}{2}) - (45° + \tfrac{\alpha}{4}) = \angle JET.$$

This is enough to conclude that $ET \perp YZ$, *i.e.* that ETI are collinear and furthermore lie on a line parallel to PJ. Thus, $JEIP$ is a parallelogram, and as long as it is not degenerate, we may conclude that $|JE| = |IP|$, *i.e.* that $\Omega_{ABC} \cong \Omega_{PQR}$.

If it is degenerate, then triangle ABC must be isosceles with A at its vertex. In this case, line MN is parallel to BC, so from $XNY \sim BPY$ we find that $|BP| = |PY|$. And since $BI \perp PY$ and $BP \perp PJ$, it follows that right triangles BPI and PYJ are congruent. Thus $|PI| = |YJ|$, and again the two circles are congruent.

Let r be the common radius length. To prove part (b), it suffices to show that $d = |IJ| = r\sqrt{3}$. Consider the inversion ι through Ω_{MNP}. The points $\iota(P)$, $\iota(Q)$, and $\iota(R)$ correspond to the midpoints of ZY, YT, and TZ respectively, so $\iota(\Omega_{ABC})$ is the nine-point-circle of triangle YTZ. In particular, the radius of $\iota(\Omega_{ABC})$ is half the radius of Ω_{MNP}, *i.e.*, $\frac{r}{2}$.

Let IJ intersect Ω_{ABC} at H and K, so that $|JH| = d - r$ and $|JK| = d + r$. Then $|\iota(HK)| = |\iota(JH)| - |\iota(JK)| = \frac{r^2}{d-r} - \frac{r^2}{d+r}$. But $\iota(HK)$ is a diameter of $\iota(\Omega_{ABC})$, and therefore has length r. This indeed gives $d = r\sqrt{3}$, as needed for part (b). □

Euler and Napoleon

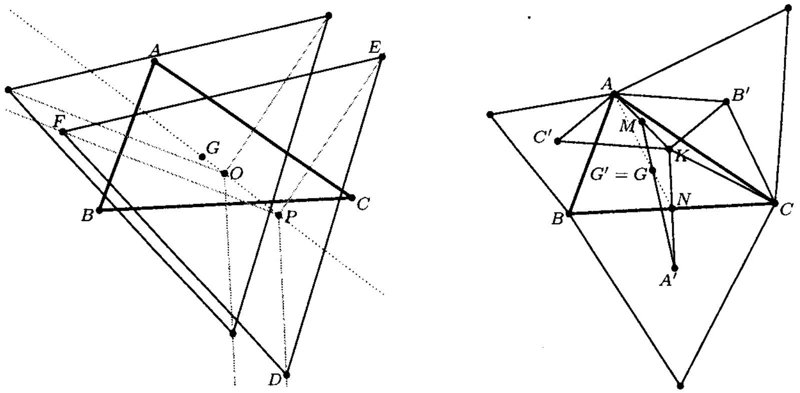
F07 – 1. Consider $\triangle ABC$ an arbitrary triangle and P a point in its plane. Let D, E, and F be three points on the lines through P perpendicular to the lines $\overline{BC}$, $\overline{CA}$, and $\overline{AB}$, respectively. Prove that if $\triangle DEF$ is equilateral and if P lies on the Euler line of $\triangle ABC$, then the center of $\triangle DEF$ also lies on the Euler line of $\triangle ABC$.

Proposed by Cosmin Pohoata (Bucharest, Romania) and Darij Grinberg (Germany).

Solution by Yasuhide Minoda (Tetsu Ryoku-Kai, Japan). Let O be the circumcenter of $\triangle ABC$. By parallel translation, it suffices to consider the case $P = O$ (see Figure 12.1(a)).

Let A', B', C' be centers of equilateral triangles outside of $\triangle ABC$, drawn on sides $\overline{BC}$, $\overline{CA}$, $\overline{AB}$, respectively. It is well known that $\triangle A'B'C'$ is equilateral (the outer Napoleon triangle). [***Editor's note:*** if $\triangle ABC$ and $\triangle DEF$ have opposite orientation, we should instead take A', B', and C' as the centers of equilateral triangles drawn *inward* on the three sides of $\triangle ABC$, so that $\triangle A'B'C'$ is the *inner* Napoleon triangle, which is also equilateral.]

Lemma 1. *$\triangle A'B'C'$ and $\triangle DEF$ are similar with respect to O. In particular, the center of $\triangle A'B'C'$, the center of $\triangle DEF$ and O are collinear.*



(a) Without loss, we may assume $P = O$.

(b) The centroids of $\triangle ABC$ and $\triangle A'B'C'$ coincide.

Figure 12.1: Figures for Problem F07 – 1.

Proof. Without loss of generality, we can assume $\angle BAC \neq 2\pi/3$. Scale up or down $\triangle DEF$ with respect to O to form $\triangle D'E'F'$ so that D' coincides with A'. We want to show $E' = B'$ and $F' = C'$.

Rotate line $\overline{OB'}$ by $\pi/3$ [or $-\pi/3$ depending on orientation] around A' and denote it by ℓ. B' and E' are moved on ℓ by this rotation. On the other hand, B' and E' are moved to C' and E', respectively, and these points are both on line $\overline{OC'}$. Thus, F' and C' must be the intersection of line $\overline{OC'}$ and ℓ (note that these lines intersect at one point because $\angle BAC \neq 2\pi/3$).

Therefore, $C' = F'$ and $B' = E'$. It follows that $\triangle A'B'C'$ and $\triangle DEF$ are similar with respect to O. □

Lemma 2. *The center of $\triangle A'B'C'$ coincides with the centroid of $\triangle ABC$.*

Proof. Let K be a point symmetric to A' with respect to line $\overline{BC}$ (see Figure 12.1(b)). Triangles $\triangle B'KC$ and $\triangle ABC$ are similar with a ratio $1 : \sqrt{3}$. In addition, $\overline{AC'} : \overline{AB} = 1 : \sqrt{3}$. Thus, $\overline{AC'} = \overline{B'K}$. Similarly, $\overline{AB'} = \overline{C'K}$. So $AC'KB'$ is a parallelogram.

Let the center of $AC'KB'$ be M, and the midpoint of $\overline{BC}$ be N. The centroid G' of $\triangle A'B'C'$ lies on $\overline{MA'}$ and $\overline{MG'} : \overline{G'A'} = 1 : 2$. Thus,

$$\frac{\overline{A'N}}{\overline{NK}} \cdot \frac{\overline{KA}}{\overline{AM}} \cdot \frac{\overline{MG'}}{\overline{G'A'}} = \frac{1}{1} \cdot \left(-\frac{2}{1}\right) \cdot \frac{1}{2} = -1.$$

So, by Menelaus' theorem, G' lies on the median $\overline{AN}$ of $\triangle ABC$. Similarly, G' also lies on other medians of $\triangle ABC$. The lemma is proved. □

By Lemma 1 and Lemma 2, the centroid of $\triangle DEF$ is on line $\overline{OG}$, namely the Euler line of $\triangle ABC$. □

Also solved by the proposers.

Dastardly Haberdashery

F07 - 2. Professor Perplex has rounded up his $n > 0$ hat-game seminar students and made the following ominous announcement:

> "I have assigned each of you a hat according to a uniform probability distribution, which I will put on your head after allowing you time to discuss a strategy. Hats come in $h > 0$ different colors, but some colors might be reused and others might not be used at all. Each student will be given a list of the h colors. Nobody will be able to see his or her own hat, but everyone will have the opportunity to observe all the other hats. Then, you will all be instructed to simultaneously write down one of the colors. If any student correctly identifies the color of his or her own hat, then there will be no final exam this semester. Otherwise, I will assign a week-long haberdashery final."

What is the probability that the students have to take a final, assuming best play?

Proposed by John Hawksley (Massachusetts Institute of Technology '08) and Scott D. Kominers '09.

Solution by Charlie Pasternak (Takoma Park Middle School). A student cannot deduce any information about his own hat color from what he sees alone. This means that the probability of a student making a correct guess, and the expected number of correct guesses over time, remains constant, so the students' best strategy is to spread out the correct guesses as thinly as possible to "waste" as few as possible.

This best strategy is: if $n \geq h$, take h students and assign them a numbers $\{0, 1, \ldots, h-1\}$, and if $n < h$, assign the students the numbers $\{0, \ldots, n-1\}$. Then, assign each color a number from 0 to $h - 1$. When the hats are put on, each student sums up the colors of the hats seen, then writes for his own hat color the color corresponding to the difference between his assigned number and the sum of the hats he sees, modulo h.

If a student's assigned number is the sum modulo h of the colors, his guess will be correct. If $n \geq h$, at least one person's assigned number is the sum modulo h of the colors, so at least one guess will be right. If $n < h$, then the chance that one of the students' assigned numbers is the sum modulo h of the colors is $\frac{n}{h}$, as is the chance of a correct guess.

Therefore, the chance of the students taking a final is $\max(0, 1 - \frac{n}{h})$. □

Also solved by Sherry Gong '11, Arnav Tripathy '11, Ray C. He (Massachusetts Institute of Technology '07), and the proposers.

Restricted Roots' Radii

F07 – 3. Find all integer monic polynomials $f(x)$ such that

(i) $f(x) = f(1 - x)$ and

(ii) all complex zeros of f lie in the disk $|z| < \sqrt[5]{2}$.

Proposed by Vesselin Dimitrov '09.

Solution by Noam D. Elkies (Harvard University). The polynomials $f_0(x) = x^2 - x$ and $f_1(x) = x^2 - x + 1$ satisfy both conditions (the latter has roots of absolute value 1 at the primitive sixth roots of unity). Therefore so does $f_0(x)^{a_0} f_1(x)^{a_1}$ for any nonnegative integers a_0 and a_1. We claim that these are the only such polynomials, indeed the only polynomials satisfying (i) whose roots lie in $|z| < r := 1.3$ (note that $2^{1/5} < 1.15 < r$).

Let f be any polynomial satisfying both conditions, and let α be a complex zero of f. Then $1 - \alpha$ is also a complex zero of f, so α lies in the intersection of the open discs of radius r about 0 and 1. We claim:

Lemma 3. *If z is a complex number such that $|z| < r$ and $|1 - z| < r$, then $|f_0(z)^2 f_1(z)^3| < 1$.*

Assuming this lemma, it follows that the algebraic integer $y := f_0(\alpha)^2 f_1(\alpha)^3$ has the property that y and all its conjugates (which are also values of $f_0^2 f_1^3$ at roots of f) have absolute value less than 1. But then the norm of y, which is the product of those conjugates, is a rational integer of absolute value less than 1. Therefore y is an algebraic number of norm zero, whence $y = 0$. This means that α is a root of either f_0 or f_1. Moreover, for each $j = 0, 1$ the two roots of f_j are switched by the involution $x \leftrightarrow 1 - x$ of $\mathbb{C}$, and thus have the same multiplicity as complex zeros of f. Letting a_j be this common multiplicity, we find that $f = f_0(x)^{a_0} f_1(x)^{a_1}$, as claimed.

We prove the lemma via the following explicit calculation. Let R be the intersection of the closed discs $|z| \leq r$ and $|1 - z| \leq r$. Let $y(z) := f_0(z)^2 f_1(z)^3$. We claim $|y(z)| \leq 1$ for all $z \in R$. The function y is analytic; hence by the maximum principle it suffices to prove $|y(z)| \leq 1$ on the boundary of R. By symmetry about the real axis and the vertical line $\text{Re}(z) = 1/2$, we may assume $z = x + i\sqrt{r^2 - x^2}$ for some $x \in [1/2, r]$. For lack of a better idea, we expand $|y(z)|^2$, obtaining a polynomial $P_8(x) \in \mathbb{Q}[x]$ of degree 8. A numerical plot suggests that this polynomial does not exceed 0.9 on $[1/2, r]$. Since $P_8(1/2) = r^8(1 - r^2)^6 < 0.9$, we can prove that $P_8(x) < 0.9$ for all $x \in [1/2, r]$ by checking that this interval contains no roots of $P_8 - 0.9$, and this is confirmed using Sturm's method (implemented in `gp` as `polsturm`). This completes the proof of the lemma and the solution of F07–3.

Remark. We could likewise use $f_0(\alpha)^{b_0} f_1(\alpha)^{b_1}$ for any positive integers b_0, b_1. The simple choice $b_0 = b_1 = 1$ suffices to solve problem F07–3 as stated, but would only let us improve r from about 1.15 to about 1.27. The choice $(b_0, b_1) = (2, 3)$ is not quite optimal either, but can be used for any r less than the positive root $r_0 = 1.304+$ of $x^{10} - 3x^8 + 3x^6 - x^4 - 1$ (for which $f_0(z)^2 f_1(z)^3 + 1$ has roots with $|z| = |1 - z| = r_0$), and numerical experimentation suggests that this r_0 is very nearly as far as this technique can be pushed. At any rate we cannot take $r_0 > 1.3503$ because the tenth-degree polynomial $f_0^2 f_1^3 + 1$ satisfies condition (i) and has a pair of complex roots of absolute value 1.35025542+.

Remark. While we would get a worse bound had we used $(b_0, b_1) = (1, 1)$ instead of the exponents $(2, 3)$ in our Lemma, the proof would be a routine albeit tedious calculus exercise, because instead of the octic P_8 we would have only a cubic polynomial to maximize. This would still be enough to solve problem F07–3 as stated, or even with the bound $2^{1/5}$ raised to $2^{1/3} = 1.2599+$. (The cutoff value r_0 would then be the positive root 1.272+ of $x^4 - x^2 - 1$.) We do not know whether the proposer's choice of $2^{1/5}$ allows for a more elegant proof. □

Editor's note. Indeed, the proposer's solution uses a similar method with $b_0 = b_1 = 1$. His choice of $\sqrt[5]{2}$ allows for a clean proof of the requisite lemma: From the identity $5(x^2 - x)(x^2 - x + 1) = x^5 + (1 - x)^5 - 1$, the inequality

$$|f_0(\alpha)f_1(\alpha)| = \frac{1}{5}\left|\alpha^5 + (1-\alpha)^5 - 1\right| < \frac{1}{5}\left(|\alpha|^5 + |1-\alpha|^5 + 1\right) = 1$$

follows immediately.

Also solved by the proposer.

A Surprisingly Constant Limit

F07 – 4. Let $a, b \geq 0$ be two nonnegative numbers. Find the limit

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{1}{n + k + b + \sqrt{n^2 + kn + a}}.$$

Proposed by Ovidiu Furdui (University of Toledo).

Solution by Paolo Perfetti (Università degli studi di Tor Vergata Roma, Math. Dept.). The key is to observe that the limit in independent of a and b. In view of this fact we compute the limit taking $a = b = 0$, getting

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{1}{n + k + \sqrt{n^2 + nk}} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \frac{1}{1 + \frac{k}{n} + \sqrt{1 + \frac{k}{n}}} = \int_0^1 \frac{1}{1 + x + \sqrt{1+x}}\, dx$$

$$= \int_1^2 \frac{1}{t + \sqrt{t}}\, dt = 2\int_1^{\sqrt{2}} \frac{1}{1+y}\, dy = 2\ln\frac{1+\sqrt{2}}{2}.$$

Now we prove that the limit is independent of a and b. First of all we write

$$\frac{1}{r + b + \sqrt{nr + a}} = \frac{1}{r + b + \sqrt{nr}\sqrt{1 + \frac{a}{nr}}} = \frac{1}{r + b + \sqrt{nr}(1 + O(\frac{1}{nr}))}$$

$$= \frac{1}{r + b + \sqrt{nr}}\left(1 + O\left(\frac{1}{nr}\right)\right)$$

and

$$\sum_{r=n+1}^{2n} \frac{O(\frac{1}{nr})}{r + b + \sqrt{nr}} = \sum_{r=n+1}^{2n} O\left(\frac{1}{nr^2}\right) = n \cdot O\left(\frac{1}{n^3}\right),$$

where the limit of the above expression as n approaches ∞ is zero. Thus,

$$\lim_{n\to\infty} \sum_{r=n+1}^{2n} \frac{1}{r + b + \sqrt{nr + a}} = \lim_{n\to\infty} \sum_{r=n+1}^{2n} \frac{1}{r + b + \sqrt{nr}}.$$

Similarly,

$$\sum_{r=n+1}^{2n} \frac{1}{r + b + \sqrt{nr}} = \sum_{r=n+1}^{2n} \frac{1}{(r + \sqrt{nr})(1 + \frac{b}{r+\sqrt{nr}})} = \sum_{r=n+1}^{2n} \frac{(1 + O(r^{-1}))}{r + \sqrt{nr}}$$

and

$$\lim_{n\to\infty}\sum_{r=n+1}^{2n}\frac{O(r^{-1})}{r+\sqrt{nr}}=\lim_{n\to\infty}\sum_{r=n+1}^{2n}O(r^{-2})=\lim_{n\to\infty}n\cdot O(n^{-2})=0,$$

so

$$\lim_{n\to\infty}\sum_{r=n+1}^{2n}\frac{1}{r+b+\sqrt{nr}}=\lim_{n\to\infty}\sum_{r=n+1}^{2n}\frac{1}{r+\sqrt{nr}}.$$

As this is independent of a and b, the proof is complete. □

Also solved by The Northwestern University Math Problem Solving Group and the proposer.

FEATURE

13

ENDPAPER

Math Has This Funny Property

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

Scott D. Kominers[†]
Harvard University '09
Cambridge, MA 02138
kominers@fas.harvard.edu

Our Mathematical Minutiae article, "i Has This Funny Property" (this *HCMR*, p. 75), details the analytic power of the complex numbers $\mathbb{C}$. Once we adjoin a single square root of -1 to the real numbers $\mathbb{R}$, we obtain startling results: Cauchy's Theorem, Liouville's Theorem, The Cauchy Integral Formula....

But if we add just a few more square roots of -1 to obtain the **quaternions** $\mathbb{H} \supset \mathbb{C}$, some of the underlying structure breaks down. In the standard presentation, the roots $\{i, j, k\}$ of -1 do not even commute:

$$ij = k = -ji, \quad jk = i = -kj, \quad ik = -j = -ki.$$

Since $\mathbb{H}$ is not commutative, it is not necessarily true that $[q(z)]^n$ is "$\mathbb{H}$-analytic" when $q(z)$ is. Thus, we figured, *everything should break down once we pass from* $\mathbb{C}$ *to* $\mathbb{H}$.

The endpaper would be called "j and k Have This Funny Property." We would give all the standard counterexamples from quaternionic analysis, proving in a tour-de-force of mathematical irony how two new roots could devolve the entire analytic system of $\mathbb{C}$ below its foundations.

We raced to the references. We read. We re-read. We—

We were wrong.

No counterexamples. In the late 1930s, Fueter obtained quaternionic analogues of Cauchy's Theorem, Liouville's Theorem, and even of power series developments. Our ironic tour-de-force was ruined before we were even born.

We despaired, until we thought to add four more square roots of -1 to obtain the **octonions** $\mathbb{O} \supset \mathbb{H} \supset \mathbb{C}$. We thought, *this extension of* $\mathbb{C}$ *is not even* associative—*there is no way octonionic analysis could be well-behaved!*

We raced to the references. We read. We—

Math is funny like that.

[†] See biographical information in this *HCMR*, p. 75.