

Bogdan Grechuk

Theorems of the 21st Century

Volume I

Theorems of the 21st Century

Bogdan Grechuk

Theorems of the 21st Century

Volume I



Springer

Bogdan Grechuk
Department of Mathematics
University of Leicester
Leicester, UK

ISBN 978-3-030-19095-8 ISBN 978-3-030-19096-5 (eBook)
<https://doi.org/10.1007/978-3-030-19096-5>

Mathematics Subject Classification (2010): 00-02, 00A05, 00A06, 00A09, 01-02, 01A61

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Most of the theorems published recently in leading mathematical journals are so difficult that even their exact formulation is hard to understand for non-mathematicians, and, in some cases, even for mathematicians working in unrelated areas. For example, in 2009 Ngô Bảo Châu proved the result called the “Fundamental lemma of the Langlands program”, which was listed by Time magazine as one of the Top 10 scientific discoveries of 2009, and earned him a Fields Medal, one of the most prestigious awards in mathematics. However, even the exact formulation of the lemma (with all the notations defined) takes several pages to write down, and requires a high-level mathematical education to understand. We can compare this result with an abstract painting or a piece of modern art, which can be fully appreciated by a relatively small group of people. However, music, poetry, paintings, films, etc., tend to be described as “best” if they can be appreciated by millions.

Do there exist mathematical theorems like the songs of The Beatles, who had stadiums of fans at their concerts? That is, theorems which are sufficiently difficult and important to be accepted for publication into a leading mathematical journal, but at the same time with a sufficiently simple formulation which can be understood and appreciated by readers with at most an undergraduate (ideally high-school) education. The *Theorems of the 21st Century* project aims to show that such theorems do exist, and possibly, that there are more such theorems than you expected.

For this book, which is the first in the *Theorems of the 21st Century* series, we examine theorems published in the first decade of the twenty-first century in the *Annals of Mathematics*, which is undoubtedly one of the leading mathematical journals, and discover that the formulation of a significant portion of them (we selected 106 theorems published between 2001 and 2010) can indeed be explained to a reader with relatively little background.

This book consists of short introductions, each being 3–4 pages in length, aimed at explaining the formulations and importance of the selected theorems. Although we sometimes refer to earlier sections of the book “for more details”, each introduction is essentially self-contained and can be read independently. Because of this,

some repetitions are unavoidable. For example, the definition of a group is repeated multiple times throughout the book.

We aim to explain each theorem to the reader with the minimal possible background. For some easier-to-understand theorems, the introductions are aimed at a high-school audience, with very little preliminary knowledge assumed: sometimes even the definition of a prime number is included. Introductions to more advanced theorems assume at least an undergraduate level.

While aiming to be accessible, we also try to maintain mathematical rigor whenever possible. Instead of adopting a newspaper-style exposition, saying that “mathematicians have proved an important result, but the details are too difficult to be presented here,” we formulate each theorem rigorously, and, when possible, give the formal definitions of all the concepts involved. The main focus is on the formulation of each theorem, its importance, and applications—the proofs of the vast majority of the theorems are not discussed at all.

Each section is devoted to one theorem, and ends with the reference to the paper in which this theorem has been proved. Some papers have several main results, in which case we have selected one of them (the most important or most accessible) as the main “theorem” to be explained, sometimes giving brief informal descriptions of the other results.

Of course, the theorems described in this book form just a small portion of the amazing mathematical discoveries of the twenty-first century. Many theorems of the highest importance, such as Perelman’s proof of the Poincaré conjecture, were not published in the *Annals of Mathematics*, while some other important theorems have been omitted because we found their formulations to be too difficult to explain. Also, the period after 2011 is not covered at all. The descriptions of other great theorems will be included in future volumes of the series.

Leicester, UK
March 2019

Bogdan Grechuk

Acknowledgements

I would like to thank Tetiana, my wife, for her continued support, encouragement, and patience while I was writing this book.

I also thank Vasyl Grechuk, my father, for many useful discussions, suggestions, and help with the figures.

I thank Paayal Mehta, Benjamin Kettley, Hiren Ladha, Sazid Balayet, and Shashank Himatrai, who did *Theorems of the 21st Century* Nuffield research projects under my supervision, and provided me with a lot of very useful feedback, suggestions, and corrections. I also thank the Nuffield Foundation for the opportunity to be a project provider with them and for finding these students for my projects.

I thank my M.Sc. student, Luke Kempsford, and my Ph.D. students, Dawei Hao and Sittichoke Som-Am, for reading the preliminary draft of this book, and for providing some suggestions and corrections.

I thank the authors of the theorems I describe in this book, prominent mathematicians who found time to read my descriptions of their theorems and provide me with very useful feedback, corrections, and suggestions. I especially thank Prof. Kevin Ford, who sent me a brilliant text about his Theorem 8.10, and allowed me to use this text in the book as is. I also received very valuable help and feedback from other authors, including Profs. Dimitris Achlioptas, Noga Alon, Marton Balazs, Itai Benjamini, Jean-Camille Birget, Yitwah Cheung, David Conlon, Jordan Ellenberg, Tamas Erdelyi, Alexandre Eremenko, Etienne Fouvry, Nicola Fusco, Loukas Grafakos, Andrew Granville, Ben Green, Roger Heath-Brown, Harald Helfgott, Bill Helton, Michael Hochman, David Hoffman, Dan Isaksen, Bo'az Klartag, Juergen Kluners, Oleg Kozlovski, Bryna Kra, Greg Kuperberg, Martin Liebeck, Mikhail Lyubich, Francesco Maggi, Jens Marklof, Christian Mauduit, William H. Meeks III, Manor Mendel, Tom Meyerovitch, Carlos Gustavo Moreira, Frank Morgan, Oleg Musin, Ken Ono, Aldo Pratelli, Omer Reingold, Joël Rivat, Jay Rosen, Muli Safra, Alexander Solynin, Jeffrey Steif, Michel Talagrand, Terence Tao, Robin Thomas, Van Vu, Matthias Weber, Michael Wolf, Trevor Wooley, Saeed Zakeri, and Ofer Zeitouni.

I thank my colleagues at the Department of Mathematics, University of Leicester, and especially Ruslan Davidchack, for reading parts of this book and providing suggestions and corrections.

I also thank the University of Leicester for granting me academic study leave, part of which I used to finish this book.

I thank Springer, and especially Rémi Lodh, for their interest in publishing this book. I also thank the referees for their suggestions, and Barnaby Sheppard for his careful copy edit of the manuscript.

Contents

1	Theorems of 2001	1
1.1	Moderate Deviations for the Volume of the Wiener Sausage	1
1.2	The Minimal Average Value of a Bounded Multiplicative Function	6
1.3	Counting Integer Solutions of Some Inequalities	10
1.4	On the Arithmetic Difference of Regular Cantor Sets	14
1.5	The Existence of Different Groups with Isomorphic Group Rings	18
1.6	Short Representations of Elements of Finite Simple Groups	23
1.7	The Existence of a Field with u -Invariant 9	26
2	Theorems of 2002	31
2.1	Counting Rational Functions with Given Critical Points	31
2.2	Representing Braids as Matrices	35
2.3	Explicit Expander Constructions Using the Zig-Zag Product	38
2.4	Elliptic Curves Over Function Fields Can Have Arbitrarily Large Rank	42
2.5	The Optimality of the Standard Double Bubble	46
2.6	Counting Integer Solutions of Equations in Three Variables	49
2.7	The Regular-Stochastic Dichotomy for Quadratic Polynomials	53
2.8	A Finitely Presented Group with an NP-Complete Word Problem	58
2.9	Finitely Generated Groups with a Word Problem in NP	63
2.10	Positive Noncommutative Polynomials are Sums of Squares	66

2.11	The Only Space Isomorphic to Each of its Subspaces	69
2.12	Counting Matrices with Some Special Properties	73
2.13	Transforming Convex Bodies into Balls	77
3	Theorems of 2003	81
3.1	On the Differentiability of Lipschitz Maps on Infinite-Dimensional Spaces	81
3.2	Representing 1 as a Sum of Reciprocals of Selected Integers	85
3.3	The Optimal Hardy–Littlewood Maximal Inequality	88
3.4	Improved Upper Bounds on Sphere Packings	92
3.5	Sums <i>Versus</i> Products of Finite Sets of Integers	95
3.6	The Radius of Integer Points in the Plane	98
3.7	The Set of Nonergodic Directions Can Have Dimension 1/2	102
3.8	A Real Number Which Is Far from All Cubic Algebraic Integers	106
4	Theorems of 2004	111
4.1	The Julia Set of Almost All Quadratic Polynomials is Locally Connected	111
4.2	The Regular Polygons have Minimal Logarithmic Capacity	116
4.3	On the Growth of the Diffusion Coefficient	119
4.4	On the Volume of the Intersection of Two Wiener Sausages	123
4.5	Controlling the Size of the Bilinear Hilbert Transform	127
4.6	Covering Convex Bodies by Balls	130
4.7	The Parametrization of Quartic Rings	133
4.8	The Time it Takes for a Random Walk to Cover the Plane	137
4.9	On the Geometry of the Uniform Spanning Forest	141
4.10	A Polynomial Time Algorithm for Primality Testing	144
5	Theorems of 2005	149
5.1	Structured Additive Patterns in Sets of Positive Density	149
5.2	A Sharp Form of Whitney’s Extension Theorem	154
5.3	Minimal Surfaces in 3-Space I	157
5.4	Statistical Properties of Quadratic Dynamics	161
5.5	Every Subset of Primes of Positive Density Contains a 3-Term Progression	164
5.6	Every Separable Infinite-Dimensional Banach Space Has Infinite Diameter	168
5.7	The NP-Hardness of the 1.36...-Approximation to the Minimum Vertex Cover	172

5.8	Embedding Large Subsets of Finite Metric Spaces into Euclidean Space	176
5.9	On the Number of Quartic Fields with Bounded Discriminant	180
5.10	Optimal Sphere Packing in Dimension 3	183
5.11	The Chromatic Number of a Random Graph	187
6	Theorems of 2006	191
6.1	Sufficient Conditions for Completeness of a Set of Integers	191
6.2	Counting Number Fields of Bounded Discriminant	194
6.3	Perfect Powers in Fibonacci and Lucas Sequences	198
6.4	A Characterization of Perfect Graphs	201
6.5	Littlewood's Conjecture Holds Outside of a Set of Dimension 0	204
6.6	The Connection Between Metric Entropy and Combinatorial Dimension	208
6.7	On the Approximation of Real Numbers by Irreducible Fractions	211
7	Theorems of 2007	217
7.1	Bounding the Error in Approximation of Smooth Functions by Polynomials	217
7.2	Superlinear Growth of Digit Patterns in the Decimal Expansion of Algebraic Numbers	221
7.3	The Existence of Intervals with Too Many and Too Few Primes	224
7.4	Interval Exchange Transformations Are Almost Always Weakly Mixing	228
7.5	The Hopf Condition Over Arbitrary Fields	231
7.6	Any Real Polynomial Can be Approximated by a Hyperbolic Polynomial	235
7.7	The Schinzel–Zassenhaus Conjecture for Polynomials with Odd Coefficients	239
7.8	Diophantine Approximation of Points on Smooth Planar Curves	243
7.9	The Hasse Principle for Systems of Two Diagonal Cubic Equations	247
7.10	An Effective Multidimensional Szemerédi Theorem	250
8	Theorems of 2008	255
8.1	Minimal Surfaces in 3-Space II	255
8.2	Arbitrarily Long Arithmetic Progressions of Prime Numbers	259
8.3	On Poincaré's Inequality	262

8.4	Representing Matrices in $SL_2(\mathbb{Z}/p\mathbb{Z})$ Using a Small Number of Generators	266
8.5	The Cayley Graphs of $SL_2(\mathbb{F}_p)$ Form a Family of Expanders	269
8.6	Determining the Shape of an Infected Region	272
8.7	A Negative Answer to Littlewood's Question About Zeros of Cosine Polynomials	276
8.8	The Kissing Number in Dimension Four is 24	279
8.9	A Criterion for Embedding L_p into L_q Uniformly	283
8.10	The Distribution of Integers with a Divisor in a Given Interval	286
8.11	An Upper Bound for the Norm of the Inverse of a Random Matrix	291
8.12	An Isoperimetric Inequality with Optimal Exponent	294
8.13	A Negative Answer to Maharam's Question	298
9	Theorems of 2009	303
9.1	On De Giorgi's Conjecture in Dimension at Most 8	303
9.2	An Efficient Algorithm for Fitting a Smooth Function to Data	307
9.3	A Helicoid-Like Surface with a Hole	310
9.4	Bounding the Condition Number of Random Discrete Matrices	314
9.5	Characterizing the Legendre Transform of Convex Analysis	318
9.6	The Solution of the Ten Martini Problem	321
9.7	A Linear Time Algorithm for Edge-Deletion Problems	325
9.8	A Characterization of Stability-Preserving Linear Operators	328
9.9	On the Gaps Between Primes	332
9.10	A Proof of the B. and M. Shapiro Conjecture in Real Algebraic Geometry	335
9.11	Bounding Diagonal Ramsey Numbers	339
9.12	An Almost Optimal Upper Bound for Moments of the Riemann Zeta Function	342
9.13	Optimal Lattice Sphere Packing in Dimension 24	346
9.14	A Waring-Type Theorem for Large Finite Simple Groups	350
10	Theorems of 2010	355
10.1	Majority Votes Are the Most Stable	355
10.2	On Divisibility Properties of Dyson's Rank Partition Function	360
10.3	Exceptional Times in Percolation Theory	364
10.4	Polynomial Parametrization for the Solutions of Diophantine Equations	367

10.5	The Order of Growth of Variance in the Asymmetric Simple Exclusion Process	370
10.6	Divergent Square Averages in Ergodic Dynamical Systems	374
10.7	Divisibility Properties of Sums of Digits of Prime Numbers	379
10.8	Prime Values of Linear Equations	383
10.9	Entropies of Multidimensional Shifts of Finite Type may be Impossible to Compute	389
10.10	Subgroups of 2-Generated Groups	393
10.11	On the Number of Quintic Fields with Bounded Discriminant	398
10.12	On the Negative Pell Equation	403
10.13	The Norms of Random Band Matrices	408
11	Further Reading	413
References		415
Author Index		431
Index		437

Acronyms

Some standard mathematical notation.

- \mathbb{Z} – the set of all integers: $\dots, -3, -2, -1, 0, 1, 2, 3, \dots$.
- \mathbb{Q} – the set of rational numbers, that is, numbers of the form $\frac{n}{m}$, where n and $m \neq 0$ are integers.
- \mathbb{R} – the set of all real numbers.
- \mathbb{R}^d – standard (Euclidean) d -dimensional space. For example, \mathbb{R}^2 is the plane, while \mathbb{R}^3 is the 3-dimensional space we live in.
- $[a, b]$ denotes the set of all real numbers x such that $a \leq x \leq b$.
- (a, b) denotes the set of all real numbers x such that $a < x < b$.
- $[a, b)$ and $(a, b]$ denote the set of all real numbers x such that $a \leq x < b$, and $a < x \leq b$, respectively.
- If x is a real number, $|x|$ denotes the absolute value of x . For example, $|3| = 3$, while $|-5| = 5$.
- $\exp(x)$ denotes e^x , where $e = 2.71828\dots$ is the base of the natural logarithm.
- \in denotes membership of a set. For example, $n \in \mathbb{Z}$ means that n is an integer, while $x \in \mathbb{R}$ states that x is a real number.
- $X \subset Y$ indicates that the set X is a subset of the set Y . For example, $\mathbb{Z} \subset \mathbb{Q}$, while $\mathbb{Q} \subset \mathbb{R}$.
- \forall – for every. For example, “ $\forall n \in \mathbb{Z} \dots$ ” means “For every integer $n \dots$ ”.
- \exists – exists. For example, “ $\exists n \in \mathbb{Z} \dots$ ” means “There exists an integer $n \dots$ ”.
- $\{\dots | \dots : \dots\}$ is a notation used to describe a set. For example, let A be the set of all even integers. In other words, A is the set of all integers n for which there exists an integer k such that $n = 2k$. This can be written as $A = \{n \in \mathbb{Z} | \exists k \in \mathbb{Z} : n = 2k\}$.
- If S is a finite set, $|S|$ denotes the number of elements in S . For example, $|\{7, 9, 11\}| = 3$.
- \sum denotes summation. For example, $\sum_{i=1}^n x_i$ is a notation for $x_1 + x_2 + \dots + x_n$. In particular, $\sum_{i=1}^3 i^2 = 1^2 + 2^2 + 3^2 = 14$.

- $\sum \sum$ denotes double summation for all pairs of the corresponding indices. For example, $\sum_{i=1}^2 \sum_{j=1}^2 x_{ij} = x_{11} + x_{12} + x_{21} + x_{22}$. Or,

$$\sum_{i=1}^2 \sum_{j=1}^2 i^2 j = 1^2 \cdot 1 + 1^2 \cdot 2 + 2^2 \cdot 1 + 2^2 \cdot 2 = 15.$$

- \prod denotes a product. For example, $\prod_{i=1}^n x_i$ is a notation for $x_1 \cdot x_2 \cdot \dots \cdot x_n$ (sometimes we omit \cdot and just write $x_1 x_2 \dots x_n$). Or,

$$\prod_{i=1}^2 \prod_{j=1}^2 (2i+j) = (2 \cdot 1 + 1)(2 \cdot 1 + 2)(2 \cdot 2 + 1)(2 \cdot 2 + 2) = 360.$$

- \cup denotes the union of sets. For example, $\{1, 2\} \cup \{2, 3, 4\} = \{1, 2, 3, 4\}$. Also, $\bigcup_{i=1}^n A_i$ denotes the union $A_1 \cup A_2 \cup \dots \cup A_n$.
- \cap denotes the intersection of sets. For example, $\{1, 2\} \cap \{2, 3, 4\} = \{2\}$. Also, $\bigcap_{i=1}^n A_i$ denotes the intersection $A_1 \cap A_2 \cap \dots \cap A_n$.
- For a positive integer n , $n!$ denotes the product of all integers from 1 to n , for example, $3! = 1 \cdot 2 \cdot 3 = 6$. We also define $0!$ to be 1.
- $:=$ denotes “equal by definition”. For example, $n! := 1 \cdot 2 \cdot \dots \cdot n$ for every positive integer n .
- \min denotes the operation of finding the minimum. For example, $\min(-2, 7, 3) = -2$. Also, $\min_{-1 \leq x \leq 2} (x^2 + 1) = 1$, because the minimum value of $x^2 + 1$ for $x \in [-1, 2]$ is equal to 1.
- Similarly, \max denotes the operation of finding maximum. For example, $\max(-2, 7, 3) = 7$, and $\max_{-1 \leq x \leq 2} (x^2 + 1) = 5$.
- \inf denotes the infimum. For a set $S \subset \mathbb{R}$, $\inf S$ is the largest real number x such that $y \geq x$ for all $y \in S$. For example, $\inf_{1 < x < 3} (x^2 + 1) = 2$.
- Similarly, \sup denotes the supremum. For a set $S \subset \mathbb{R}$, $\sup S$ is the smallest real number x such that $y \leq x$ for all $y \in S$. For example, $\sup_{1 < x < 3} (x^2 + 1) = 10$.
- $a \equiv b \pmod{c}$ means that integers a and b give the same remainder after division by c , or, in other words, $a - b$ is divisible by c . For example, $11 \equiv 2 \pmod{3}$.

Chapter 1

Theorems of 2001



1.1 Moderate Deviations for the Volume of the Wiener Sausage

Heat Conduction, and Particle Trajectories

When you turn on a radiator in your house, it first heats the air within a small distance from it. How long will it take for the “hot air” to “spread out”, mix well with the “cold air”, and make your house uniformly warm? This process is called *heat conduction*. To understand it, let us look at the trajectory of an individual particle which starts near the radiator. Of course, this trajectory is complicated, frequently changing direction (due to particle collisions), but initially it is *local*, that is, it moves within a small region near the radiator. If we wait long enough, however, the particle will eventually travel throughout the room. To understand heat conduction (and other similar important processes and phenomena), we need to understand how “big” the region “covered” by the trajectory of a particle gets before any given time t .

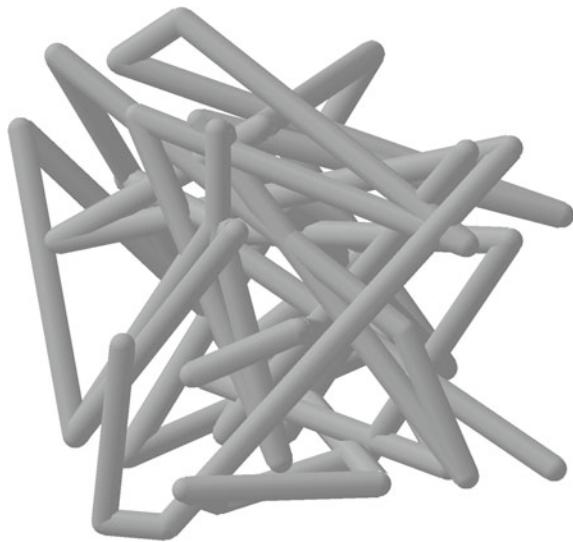
The Volume of the “Covered” Space

Let us be a bit more formal. Particle trajectories can be studied in any dimension d : for example, for $d = 1$, we can study particles moving in a very thin tube. Let the particle be modelled as a point, and let $W(t) = (w_1(t), \dots, w_d(t))$ denotes its coordinates at time t . For any $a > 0$, define

$$W^a(t) = \{x \in \mathbb{R}^d \mid \exists s \in [0, t] : \rho(x, W(s)) \leq a\}, \quad (1.1)$$

where ρ denotes the usual distance in \mathbb{R}^d . In other words, $W^a(t)$ is the set of all points at distance at most a from the particle trajectory (Fig. 1.1). The question

Fig. 1.1 The set $W^a(t)$ in (1.1) for a piecewise-linear curve $W(t)$ in dimension $d = 3$



“How big is the region covered by the trajectory?” can be formalized as “What is the (d -dimensional)¹ volume of $W^a(t)$?”. We will denote this volume by $|W^a(t)|$.

For example, let us assume for a moment that $d = 3$ and the particle moves in a straight line with constant velocity v during the time period t . In this case, $W(t)$ is a line segment of length vt , and $W^a(t)$ is a cylinder with radius a , height vt , and with semi-spheres attached to the top and the bottom. Hence, its total volume is $|W^a(t)| = \frac{4}{3}\pi a^3 + \pi a^2 vt$. In particular, $|W^a(t)|$ grows as a linear function of t . One could derive a similar result if $W(t)$ is a non-straight but smooth curve without too many self-intersections, because in this case $W^a(t)$ is essentially the same cylinder but bent.

However, the trajectories of the movement of actual particles are very far from being either straight or smooth lines. In fact, individual particle trajectories are so complicated and unpredictable that the best we can do is to consider them as “random”, and study them using the language and tools of *probability theory*. That is, instead of answering questions like “Where will this particle be after time t ?”, we will be talking about *the chance* that the particle, after time t , will be in this or that region.

A Simple Example of Random Movement

To understand what we mean by a “random trajectory” imagine a very drunk man moving along a street. He makes one step every second, but the direction of every

¹For example, for $d = 3$ we are talking about the “usual” volume, for $d = 2$ about the covered area, while for $d = 1$, about the length of the covered interval.

step may be left or right, with equal chance. For example, after 2 steps, he can move right and then again right (we denote this scenario as *RR*), or right and then left (*RL*), or left and then right (*LR*), or left and again left (*LL*). If we put him on the coordinate line, assume that he starts at 0, and his step length is 1, then, after 2 steps, he will reach point 2 in the *RR* scenario, return to 0 in the *LR* and *RL* scenarios, and reach point -2 in the *LL* scenario. In other words, if X_t denotes his position after t steps, then X_2 may be equal to 2, 0, or -2 . We cannot answer the question “what is X_2 ”, we can only list its possible values, and we can also talk about the chances, or *probabilities*, that these values will happen. In our case, the probability that $X_2 = 2$, denoted as $P(X_2 = 2)$, is equal to $1/4$, because this can happen in 1 out of 4 possible scenarios. Similarly, $P(X_2 = 0) = 2/4 = 1/2$, because $X_2 = 0$ in 2 scenarios (*LR* and *RL*) out of 4. Finally, $P(X_2 = -2) = 1/4$.

The Average Length of the Covered Interval

Now, let V_t be the length of the interval the man “covered” up to time t . For example, let $t = 2$. In the *RR* scenario, the man moves straight from 0 to 2, the covered interval is $[0, 2]$, and $V_2 = 2$. In the *RL* scenario, the man moved from 0 to 1 and back to 0, hence he covered only the interval $[0, 1]$, and $V_2 = 1$. Similarly, in the *LR* scenario, the covered interval is $[-1, 0]$, and $V_2 = 1$, while in the *LL* scenario, the covered interval is $[-2, 0]$, and $V_2 = 2$. In summary, V_2 can take values 1 or 2, with probabilities $P(V_2 = 1) = P(V_2 = 2) = 2/4 = 1/2$. If $t = 3$, there are 8 scenarios of movements, *RRR*, *RRL*, *RLR*, *RLL*, *LRR*, *LRL*, *LLR*, and *LLL*, and a similar calculation shows that V_3 can take values 1, 2, and 3, with probabilities $P(V_3 = 1) = P(V_3 = 3) = 2/8 = 1/4$, and $P(V_3 = 2) = 4/8 = 1/2$.

In general, after t steps, there are $m = 2^t$ possible scenarios of movements, which we can denote as $\omega_1, \dots, \omega_m$ ($\omega_1 = \text{RRR} \dots \text{RR}$, $\omega_2 = \text{RRR} \dots \text{RL}$, \dots , $\omega_m = \text{LLL} \dots \text{LL}$). For each ω_i , $i = 1, 2, \dots, m$, we can calculate the length $V_t(\omega_i)$ of the covered interval in this scenario, which then allow us to list all possible values of V_t with the corresponding probabilities.

While we cannot predict what the actual value of V_t will be (because we do not know which particular scenario the drunk man will implement), we can at least calculate the *average* (also called *expected*) value of V_t over all possible scenarios: $E[V_t] = \frac{1}{m} \sum_{i=1}^m V_t(\omega_i)$. Equivalently, $E[V_t] = \sum_{i=1}^k v_i \cdot p_i$, where v_1, v_2, \dots, v_k is the list of values V_t may take, and p_1, p_2, \dots, p_k are the corresponding probabilities. For example, $E[V_2] = \frac{1}{4}(1 + 1 + 2 + 2) = 1 \cdot \frac{2}{4} + 2 \cdot \frac{2}{4} = 1.5$, while $E[V_3] = \frac{1}{8}(1 + 1 + 2 + 2 + 2 + 2 + 3 + 3) = 1 \cdot \frac{2}{8} + 2 \cdot \frac{4}{8} + 3 \cdot \frac{2}{8} = 2$.

The Probabilities of “Moderate” Deviations from the Average

The value of $E[V_t]$ gives us an approximate idea about how large V_t is. However, how good is this “approximation”? For example, what is the probability that the actual

covered length V_t will be no greater than, say, 70% of its average value $E[V_t]$? For $t = 2$, $P(V_2 \leq 0.7E[V_2]) = P(V_2 \leq 0.7 \cdot 1.5) = P(V_2 \leq 1.05) = 2/4 = 1/2$. For $t = 3$, $P(V_3 \leq 0.7E[V_3]) = P(V_3 \leq 0.7 \cdot 2) = P(V_3 \leq 1.2) = 2/8 = 1/4 < 1/2$. If we could prove that, for large t , $P(V_t \leq 0.7E[V_t])$ becomes negligibly small, and the same is true for $P(V_t \geq 1.3E[V_t])$, we could conclude that, with very high probability, $0.7E[V_t] < V_t < 1.3E[V_t]$, that is, $E[V_t]$ approximates V_t with at least 70% accuracy. Similarly, proving that $P(V_t \leq 0.99E[V_t])$ and $P(V_t \geq 1.01E[V_t])$ are small for large t would allow us to conclude that $E[V_t]$ approximates V_t with at least 99% accuracy. Probabilities of the form $P(V_t \leq cE[V_t])$ for some constant $c < 1$ are called probabilities of *moderate*² deviations. Estimates from above for such probabilities are crucial for understanding how well $E[V_t]$ approximates V_t .

The Standard Model of Particle Movement: The Wiener Process

In some sense, particle movement resembles the movement of a drunk man considered above. If the particle is modelled as a point and forces are ignored, we can assume that it moves in a straight line until the first collision, then changes direction randomly, moves until the next collision, changes direction again, and so on.

The standard model of particle movement is called the *Wiener process*. If $d = 1$ (movement in a thin tube), the Wiener process is, intuitively, the limiting case of the movement of the drunk man. That is, we assume that the step length of the man is ε , the number of steps is $N = 1/\varepsilon^2$ per unit of time, and then let ε go to 0. The intuition in dimensions $d \geq 2$ is similar.

The Volume of a Wiener Sausage

If $W(\cdot)$ is a Wiener process, then $W^a(t)$, as defined in (1.1), is called a *Wiener sausage*. It was introduced in 1964 by Frank Spitzer [355], and since then has been used in the description of a number of physical phenomena, including heat conduction. It is known [355] (but the proof is too difficult to be presented here) that the average (or expected) volume $E|W^a(t)|$ is, for large t , approximately equal to

$$E|W^a(t)| \approx \begin{cases} \sqrt{8t/\pi}, & \text{if } d = 1, \\ 2\pi t / \ln t, & \text{if } d = 2, \\ 2\pi at, & \text{if } d = 3, \end{cases} \quad (1.2)$$

and similarly $E|W^a(t)| = C(a, d)t$ for $d \geq 3$, where $C(a, d)$ is a constant depending on a and d .

²The term “moderate” comes from the fact that probabilities of the form $P(V_t \leq f(t)E[V_t])$ for some function $f(t)$ such that $\lim_{t \rightarrow \infty} f(t) = 0$ are known as probabilities of *large* deviations.

To understand how close $|W^a(t)|$ is to its average value (1.2), we need to have a good estimate from above for the probability of moderate deviation $P(|W^a(t)| \leq cE|W^a(t)|)$ for $c < 1$. The following theorem of M. van den Berg, E. Bolthausen and F. den Hollander [384] answers this question in all dimensions $d \geq 2$.

Theorem 1.1 *For every $a > 0$, $b \in (0, 2\pi)$,*

$$\lim_{t \rightarrow \infty} \frac{1}{\ln t} \ln P(|W^a(t)| \leq bt / \ln t) = -I^{2\pi}(b), \quad d = 2,$$

and for every $a > 0$, $b \in (0, C(a, d))$,

$$\lim_{t \rightarrow \infty} \frac{1}{t^{(d-2)/d}} \ln P(|W^a(t)| \leq bt) = -I^{C(a,d)}(b), \quad d \geq 3.$$

Here, $C(a, d)$ is the constant defined after Eq. (1.2), and $I^{C(a,d)}(b)$ is the “rate function”, for which a detailed analysis is given, with some analytic formulas, estimates, graphs, etc. In short, Theorem 1.1 completely resolves the problem of estimating the probabilities in question for large t .

Some Special Cases

For $d = 2$, $b = 1.98\pi$, and large t , Theorem 1.1 implies that

$$P(|W^a(t)| \leq 1.98\pi t / \ln t) \approx t^{-C},$$

where $C = I^{2\pi}(1.98\pi)$ is a positive constant. So, the probability that volume $|W^a(t)|$ is just 1% less than the average value $2\pi t / \ln t$ goes to 0 as $t \rightarrow \infty$.

Similarly, for $d = 3$, $b = 1.98\pi a$, and large t ,

$$P(|W^a(t)| \leq 1.98\pi at) \approx \exp(-Ct^{1/3}),$$

for some $C > 0$. In this case, the probability of 1% volume deviation from average not only goes to 0, but in fact decreases exponentially fast with t . That is, if we wait a little bit, we are guaranteed that the trajectory volume of essentially all particles will be within 1% from $2\pi at$. Obviously, 1% here can be replaced by any other arbitrary small constant.

Reference

M. van den Berg, E. Bolthausen and F. den Hollander, Moderate deviations for the volume of the Wiener sausage, *Annals of Mathematics* **153**-2, (2001), 355–406.

1.2 The Minimal Average Value of a Bounded Multiplicative Function

Multiplicative Functions with Small Average Values

A function $f : \mathbb{N} \rightarrow \mathbb{R}$, where \mathbb{N} is the set of positive integers and \mathbb{R} is the real line, is called *completely multiplicative* if $f(mn) = f(m)f(n)$ for all positive integers m, n . Assuming that $|f(n)| \leq 1$ for all $n \in \mathbb{N}$, what can the average value $\frac{1}{x} \sum_{n \leq x} f(n)$ be for large x ? Because $|f(n)| \leq 1$ for all n ,

$$\left| \frac{1}{x} \sum_{n \leq x} f(n) \right| \leq \frac{1}{x} \sum_{n \leq x} |f(n)| \leq \frac{1}{x} \sum_{n \leq x} 1 \leq \frac{1}{x} \cdot x = 1,$$

hence $\frac{1}{x} \sum_{n \leq x} f(n)$ is always between -1 and 1 . For the function $f(n) = 1, \forall n$, this average is equal to 1 , the maximal possible. However, it is not clear whether the lower bound -1 is achievable. The average is -1 for the function $f(n) = -1, \forall n$, but it is not completely multiplicative (for example, $f(6) = -1 \neq 1 = f(2)f(3)$).

Let us try to ensure that the values $f(n)$ are as small as possible. Because $f(1) \cdot f(1) = f(1 \cdot 1) = f(1)$, we have $f(1) = 0$ or $f(1) = 1$. If $f(1) = 0$, then for any n , $f(n) = f(n \cdot 1) = f(n) \cdot f(1) = f(n) \cdot 0 = 0$, and $\frac{1}{x} \sum_{n \leq x} f(n) = 0$. To try to do better, we should choose $f(1) = 1$. Following our strategy to assign values as small as possible, we can let $f(2) = f(3) = -1$, but then $f(4) = f(2)f(2) = 1$. Next, we can assign $f(5) = -1$, but then $f(6) = f(2)f(3) = 1$. Continuing, we can choose $f(7) = -1$, and also $f(8) = f(4)f(2) = -1$, but then $f(9) = f(3)f(3) = 1$ and $f(10) = f(2)f(5) = 1$. Checking the average so far, we get $\frac{1}{10} \sum_{n \leq 10} f(n) = 0$: no progress!

The Proof of a Conjecture of Hall

In 1996, Richard Hall [187] constructed an example of a completely multiplicative function f with average value $\frac{1}{x} \sum_{n \leq x} f(n)$ for large x approximately equal to

$$\delta_1 = 1 - 2 \ln(1 + \sqrt{e}) + 4 \int_1^{\sqrt{e}} \frac{\ln t}{t+1} dt \approx -0.656999, \quad (1.3)$$

and conjectured that this is the lowest possible. This conjecture was proved by A. Granville and K. Soundararajan [172].

Theorem 1.2 *For a completely multiplicative function f taking values in $[-1, 1]$, we have*

$$\delta_1 \leq \lim_{x \rightarrow \infty} \frac{1}{x} \sum_{n \leq x} f(n) \leq 1, \quad (1.4)$$

where δ_1 is given by (1.3). Conversely, for any $\delta \in [\delta_1, 1]$ there exists an f as above such that the limit is equal to δ .

Quadratic Residues and Non-residues

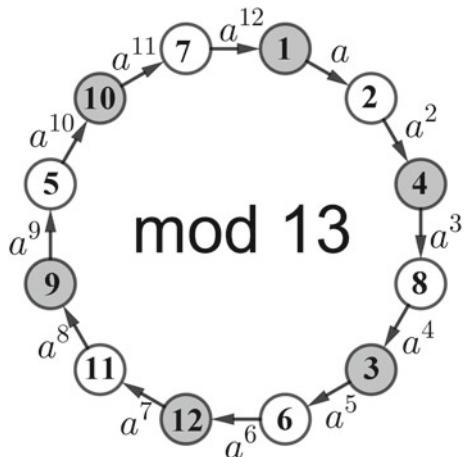
Applying Theorem 1.2 to various functions f , we can derive a number of interesting and non-trivial results. Here is one example. An integer n is called a *quadratic residue* modulo an integer p if $x^2 - n$ is divisible by p for some integer x , and a *quadratic non-residue* otherwise. For example, $2^2 - 1$ is divisible by 3, hence $n = 1$ is a quadratic residue modulo $p = 3$. However, what about $n = 2$? Can $x^2 - 2$ be divisible by 3? In fact, no. Indeed, every integer x can be written as either $x = 3k$, or $x = 3k + 1$, or $x = 3k + 2$, for some integer k . In the first case, $x^2 - 2 = 9k^2 - 2$ is not divisible by 3. Similarly, in the second case $x^2 - 2 = (3k + 1)^2 - 2 = 9k^2 + 6k + 1 - 2 = 3(3k^2 + 2k) - 1$, while in the third case $x^2 - 2 = (3k + 2)^2 - 2 = 9k^2 + 12k + 4 - 2 = 3(3k^2 + 4k) + 2$. In any case, $x^2 - 2$ is not divisible by 3, hence $n = 2$ is not a quadratic residue modulo $p = 3$.

In general, to check if $x^2 - n$ is divisible by p for some x , we should consider p cases: $x = pk, x = pk + 1, \dots, x = pk + (p - 1)$. For case $x = pk + r$, $x^2 - n = (pk + r)^2 - n = p(pk^2 + 2kr) + r^2 - n$, hence $x^2 - n$ is divisible by p if and only if $r^2 - n$ is. For simplicity, assume that $0 < n < p$. Then n is a quadratic residue modulo p if and only if r^2 gives remainder n after division by p for some $r = 1, 2, \dots, p - 1$. For example, for $p = 5$, numbers $1^2, 2^2, 3^2$ and 4^2 give remainders 1, 4, 4, and 1, respectively, after division by 5, hence 1 and 4 are quadratic residues modulo $p = 5$, while 2 and 3 are not. Similarly, for $p = 7$, $1^2, 2^2, 3^2, 4^2, 5^2$ and 6^2 give remainders 1, 4, 2, 2, 4 and 1, respectively, hence 1, 2 and 4 are quadratic residues, while 3, 5 and 6 are quadratic non-residues.

The Number of Quadratic Residues

In all these examples, exactly half of the positive integers less than p are quadratic residues, and half are non-residues. This is not a coincidence: in fact, this half-half distribution is true for every odd prime p . This is because for every odd p there exists an a such that the integers $a, a^2, a^3, \dots, a^{p-1}$ all give different remainders modulo p . Then the remainders corresponding to a^2, a^4, \dots, a^{p-1} are quadratic residues, while the ones corresponding to a, a^3, \dots, a^{p-2} are quadratic non-residues. Figure 1.2 illustrates this fact for $p = 13$ and $a = 2$.

Fig. 1.2 Quadratic residues and non-residues modulo $p = 13$



Quadratic residues have been intensively studied since the 17th and 18th centuries, but some basic questions about their distribution proved to be very difficult. For example, for some $p > 200$, can it be that all numbers from 1 to 100 are quadratic non-residues? Ok, they cannot, because perfect squares 1, 4, 9, 16, 25, 36, 49, 64, 100 are obviously quadratic residues, but can it be that all other 90 numbers are quadratic non-residues? More generally, for any large number x , can we find $p > x$ such that 90% of all numbers less than x are quadratic non-residues modulo p ? If this is impossible, what is the highest percentage of non-residues up to x we can achieve?

The 17.15% Law

Theorem 1.2 can be used to answer this difficult question in a few lines. For an odd prime p , define

$$f_p(n) = \begin{cases} 0, & \text{if } n \text{ is divisible by } p \\ 1, & \text{if } n \text{ is a quadratic residue modulo } p, \text{ but not divisible by } p, \\ -1, & \text{if } n \text{ is not a quadratic residue modulo } p. \end{cases}$$

One can check that $f_p(n)$ is a completely multiplicative function. For example, if $f_p(n) = 0$, then n is divisible by p , hence nm is divisible by p for every m , and $f_p(nm) = 0 = f_p(n) \cdot f_p(m)$. As a different example, consider the case $f_p(n) = f_p(m) = 1$, that is, both n and m are quadratic residues. Then $x^2 - n$ is divisible by p for some x , hence $x^2 = ap + n$ for some integer a . Similarly, $y^2 = bp + m$ for some integers y and b . Then $(xy)^2 = (ap + n)(bp + m) = p(abp + am + bn) + nm$, hence $(xy)^2 - nm$ is divisible by p , and, by definition, nm is a quadratic residue

modulo p , or $f_p(nm) = 1 = f_p(n) \cdot f_p(m)$. The proofs of $f_p(nm) = f_p(n) \cdot f_p(m)$ in the other cases are just a bit more complicated.

For $0 < n < p$, the expression $\frac{1+f_p(n)}{2}$ is equal to 1 if n is a quadratic residue modulo p and 0 otherwise, so the number of quadratic residues modulo p up to any $x < p$ is exactly equal to the sum $\frac{1}{2} \sum_{n \leq x} (1 + f_p(n))$.

Now, applying Theorem 1.2 to $f_p(n)$, we get

$$\lim_{x \rightarrow \infty} \frac{1}{x} \sum_{n \leq x} \frac{1 + f_p(n)}{2} \geq \frac{1 + \delta_1}{2} \approx 0.171500.$$

In other words, we have proved the following statement: *If x is sufficiently large then, for all primes $p > x$, more than 17.15% of the integers up to x are quadratic residues modulo p .* This statement also holds for $p \leq x$, and the estimate is the best possible.

Similarly, for any power $m > 2$, we can say that n is an m -th power residue modulo p if $x^m - n$ is divisible by p for some integer x . In this case, Granville and Soundararajan proved a similar result: for a given integer $m > 2$, there exists a constant $\pi_m > 0$ such that, if x is sufficiently large, then, for all primes p , more than $\pi_m\%$ of the integers up to x are m -th power residues modulo p .

Extension to Complex-Valued Functions

Granville and Soundararajan also extended Theorem 1.2 to complex-valued functions, which led to much more interesting results and applications. The set \mathbb{C} of complex numbers consists of numbers of the form $z = x + y\sqrt{-1}$, where x, y are real numbers. Any complex number can be represented as a point (x, y) in the coordinate plane. The absolute value of a complex number z is $|z| = \sqrt{x^2 + y^2}$. Let S be any set of complex numbers such that $|z| \leq 1$ for any $z \in S$. Geometrically, S is a subset of the unit disk $U = \{(x, y) \mid x^2 + y^2 \leq 1\}$. Let $F(S)$ be the set of all completely multiplicative functions $f : \mathbb{N} \rightarrow \mathbb{C}$ such that $f(p) \in S$ for any prime p . Then we can define $\Gamma_N(S)$ to be the set of complex numbers z representable in the form $z = \frac{1}{N} \sum_{n \leq N} f(n)$ for some $f \in F(S)$, and $\Gamma(S) = \lim_{N \rightarrow \infty} \Gamma_N(S)$.

Granville and Soundararajan called $\Gamma(S)$ the *spectrum* of the set S . In this notation, Theorem 1.2 corresponds to the special case $S = [-1, 1]$, and can be formulated as $\Gamma([-1, 1]) = [\delta_1, 1]$. However, their theory goes far beyond this special case—they proved many interesting properties of $\Gamma(S)$ for a general set S . For example, they proved that $\Gamma(S)$, when drawn in the coordinate plane, always looks like a connected picture, not a collection of disconnected pieces. However, an exact formula for $\Gamma(S)$ has been obtained only in the case $S = [-1, 1]$, and its extension to general S remains an intriguing open question.

Reference

A. Granville and K. Soundararajan, The spectrum of multiplicative functions, *Annals of Mathematics* **153**-2, (2001), 407–470.

1.3 Counting Integer Solutions of Some Inequalities

Counting Integer Points in Disks and Balls

How many pairs of integers (x, y) are solutions to the inequality $x^2 + y^2 \leq 100$? We can answer this question approximately by first describing its *real* solutions. In the coordinate plane, real solutions to this inequality form a disk with center $(0, 0)$ and radius $r = 10$. The question is, how many points with integer coefficients does this disk contain? Let us put into correspondence to every such point (x, y) a unit square, with vertices $(x, y), (x, y+1), (x+1, y+1), (x+1, y)$. If (x, y) is not close to the boundary of the circle, this square lies fully within it. Hence, the number of integer solutions to $x^2 + y^2 \leq 100$ is approximately the number of unit squares within the circle, which, in turn, is approximately equal to its area, see Fig. 1.3. The latter can be easily computed, and is equal to $\pi r^2 = 100\pi \approx 314$. In fact, the exact number of integer solutions to $x^2 + y^2 \leq 100$ is 309.

Similarly, the number of integer solutions to the inequality $x^2 + y^2 + z^2 \leq 100$ can be approximated by the volume of the corresponding ball, which is equal to $\frac{4}{3}\pi r^3 = \frac{4}{3}\pi 10^3 \approx 4189$. For the more general equation, $x_1^2 + x_2^2 + \dots + x_n^2 \leq m$, we need to calculate the volume of the n -dimensional ball of radius $r = \sqrt{m}$. What is the formula for it? Well, such a ball is just a ball with radius 1 enlarged by a factor of r . If we take any figure (not necessary a ball) in n -dimensional space, and enlarge it by a factor of r , its volume increases by a factor of r^n . Hence, the volume of the n -dimensional ball of radius r is Vr^n , where V is the volume of the ball $x_1^2 + x_2^2 + \dots + x_n^2 \leq 1$. Because $r = \sqrt{m}$, the volume is $V(\sqrt{m})^n = Vm^{n/2}$.

Monomials, Polynomials, and Some Volume Estimates

In general, a *monomial* in n variables x_1, x_2, \dots, x_n is any expression of the form $G(x_1, x_2, \dots, x_n) = x_1^{a_1} x_2^{a_2} \dots x_n^{a_n}$ for some non-negative integers a_1, a_2, \dots, a_n . The *degree* d of the monomial is just $a_1 + a_2 + \dots + a_n$. For example, $G(x, y, z) = xy^3z^2$ is a monomial of degree $d = 1 + 3 + 2 = 6$. If G is any monomial of degree d , then, for any $k \in \mathbb{R}$,

$$\begin{aligned} G(kx_1, kx_2, \dots, kx_n) &= (kx_1)^{a_1}(kx_2)^{a_2} \dots (kx_n)^{a_n} \\ &= k^d \cdot (x_1^{a_1} x_2^{a_2} \dots x_n^{a_n}) \\ &= k^d G(x_1, x_2, \dots, x_n). \end{aligned}$$

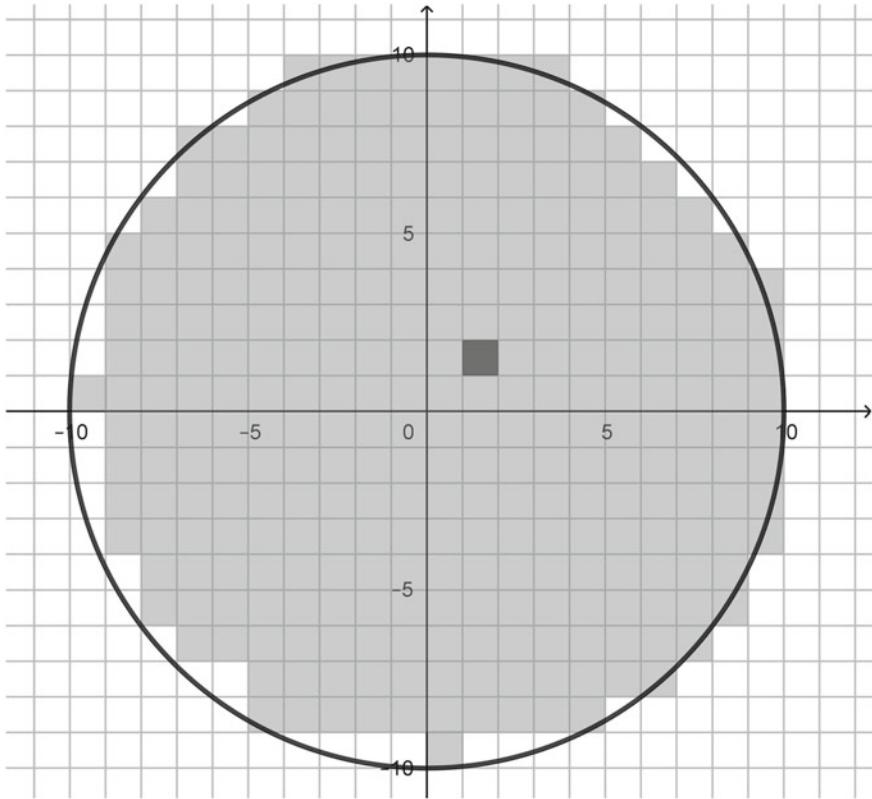


Fig. 1.3 Counting integer points in a disk

For example, if $G(x, y, z) = xy^3z^2$, then we have $G(kx, ky, kz) = (kx)(ky)^3(kz)^2 = k^6G(x, y, z)$.

A *polynomial* in n variables x_1, x_2, \dots, x_n is any sum of monomials, for example, $xy + xy^3z^2 + z^{15}$. In this example, we have a sum of three monomials of degree 2, 6, and 15, respectively. Here, we study only polynomials F which are the sums of monomials of the same degree. In particular, $F(x, y, z) = x^6 + yz^5 + xy^3z^2$ is an example of such a polynomial in $n = 3$ variables of degree $d = 6$. In this example, for any $k \in \mathbb{R}$, we have

$$\begin{aligned} F(kx, ky, kz) &= (kx)^6 + (ky)(kz)^5 + (kx)(ky)^3(kz)^2 \\ &= k^6(x^6 + yz^5 + xy^3z^2) \\ &= k^6F(x, y, z). \end{aligned}$$

In general, we have

$$F(kx_1, \dots, kx_n) = k^d F(x_1, \dots, x_n). \quad (1.5)$$

Now, let F be any polynomial in n variables with integer coefficients such that every monomial has degree d , and let the volume of the set $S_1 \subset \mathbb{R}^n$ defined by the inequality $|F(x_1, \dots, x_n)| \leq 1$ be equal to $V(F)$. Then what is the volume of the set $S_m \subset \mathbb{R}^n$ defined by $|F(x_1, \dots, x_n)| \leq m$? It follows from (1.5) that S_1 enlarged by a factor of k is defined by the equation $|F(x_1, \dots, x_n)| \leq k^d$, hence S_m is just S_1 enlarged by a factor of $k = m^{1/d}$. Thus, the volume of S_m is $V(F)k^n = V(F)m^{n/d}$.

Inequalities of Finite Type

Let $N_F(m)$ be the number $N_F(m)$ of integer solutions to $|F(x_1, \dots, x_n)| \leq m$. Motivated by the examples above, we might hope that $N_F(m)$ is approximately equal to the volume of S_m , that is,

$$N_F(m) \approx V(F)m^{n/d}. \quad (1.6)$$

Unfortunately, this is not always the case. For example, real solutions to $|x - y| \leq 0$ form a line $y = x$, and the 2-dimensional area of a line is 0. However, the number of integer solutions is obviously infinite. We say that an inequality of the form $|F(x, y)| \leq m$ is of *finite type* if the area of its set of real solutions is finite, and if its intersection with any line with rational coefficients has finite length. Similarly, if S is a set of real solutions to $|F(x, y, z)| \leq m$, then F is of finite type if the 3-dimensional volume of S is finite, the intersection of S with any plane with rational coefficients has finite area, and its intersection with any line with rational coefficients has finite length. The same definition extends to any dimension.

Decomposable Forms

Now, if the inequality $|F(x_1, \dots, x_n)| \leq m$ is of finite type, does it mean that it has a finite number of solutions, and can this number be bounded in terms of the n -dimensional volume of the set of its real solutions? In general, no: take $F(x, y) = 0$ if $y = x^2$ and $F(x, y) > m$ otherwise. Then the real solutions form a parabola $y = x^2$, it has area 0, and it intersects any line in at most two points, hence it is of finite type, but the number of integer solutions is infinite. However, Jeffrey Lin Thunder [376] proved that the answer to the above question is “yes” for functions F called *decomposable forms*. These are polynomials of degree d in n variables which are expressible as

$$F(x_1, \dots, x_n) = (a_{11}x_1 + \dots + a_{1n}x_n)(a_{21}x_1 + \dots + a_{2n}x_n) \dots (a_{d1}x_1 + \dots + a_{dn}x_n),$$

where the coefficients a_{ij} are non-zero complex numbers, that is, numbers of the form $x + y\sqrt{-1}$, with x, y being real numbers. For example, $F = x^2 + y^2$ belongs to this class, because $x^2 + y^2 = (x + y\sqrt{-1})(x - y\sqrt{-1})$.

Theorem 1.3 *Let F be a decomposable form of degree d in n variables with integer coefficients. Then the number $N_F(m)$ of integer solutions to the inequality $|F(x_1, \dots, x_n)| \leq m$ is finite for all m if and only if F is offnite type. Moreover, if F is offnite type, then $N_F(m) \leq c(n, d)m^{n/d}$, where $c(n, d)$ is an effectively computable constant depending only on n and d .*

Mahler [258] obtained similar results for $d = 2$ in 1933, and then essentially no progress was made in the general case $d > 2$ for almost 70 years, until the work of Thunder. Note that the bound in Theorem 1.3 does not depend on the coefficients of F . That is, if we fix n, d, m and compute that $c(n, d)m^{n/d}$ is, say, one million, then we can be sure that for any (integer) coefficients of a decomposable form F of degree d in n variables, the inequality $|F(x_1, \dots, x_n)| \leq m$ either has an infinite number of solutions, or at most a million. This resembles the fact that a quadratic equation $ax^2 + bx + c = 0$ can have either an infinite number of real solutions (if $a = b = c = 0$), or at most 2, but never exactly 3.

Thunder also proved that, under the conditions of Theorem 1.3, the approximation (1.6) works well for large m . Intuitively, the reason why m should be large is to remove some ‘‘boundary effects’’. For example, the inequality $x^2 + y^2 \leq m$ with $m = 0.99$ has just 1 integer solution, $x = y = 0$, while the corresponding volume is $0.99\pi > 3$. With $m = 1$, the volume is still just above 3, but the number of solutions jumps to 5. For large m , such effects are negligible, and (1.6) works. Thunder also derived the exact form of the error term in this approximation.

A Concrete Example

As an application, let us approximately count the number of integer solutions to the inequality $x^4 + 4y^4 \leq 10^{10}$. First, let us factor this polynomial to check that it is a decomposable form.

$$x^4 + 4y^4 = (x^2)^2 - (2y^2i)^2 = (x^2 - 2y^2i)(x^2 + 2y^2i),$$

where $i = \sqrt{-1}$. Note that $(1+i)^2 = 1^2 + 2i + i^2 = 2i$, hence $x^2 - 2iy^2 = (x - (1+i)y)(x + (1+i)y)$. Similarly, $(i-1)^2 = -2i$, and $x^2 + 2y^2i = x^2 - ((i-1)y)^2 = (x - (i-1)y)(x + (i-1)y)$. Hence,

$$x^4 + 4y^4 = (x - (1+i)y)(x + (1+i)y)(x - (i-1)y)(x + (i-1)y),$$

as required. Next, we need to calculate the area V of the shape $x^4 + 4y^4 \leq 1$. By symmetry, this is twice the area below the curve $y = \sqrt[4]{\frac{1-x^4}{4}}$ for $-1 \leq x \leq 1$, which can be found by integration:

$$V = 2 \int_{-1}^1 \sqrt[4]{\frac{1-x^4}{4}} \approx 1.311.$$

Finally, we apply (1.6) to conclude that the number of integer solutions in question is approximately

$$N \approx Vm^{n/d} \approx 1.311(10^{10})^{2/4} = 131100.$$

The method itself is based on the volume intuition and was known long before Theorem 1.3 was proved, always giving an accurate result in practice. However, there was no formal *proof* that this method *should* work. The work of Thunder finally filled this gap, and now the method above can be applied with full confidence.

Reference

J.L. Thunder, Decomposable form inequalities, *Annals of Mathematics* 153-3, (2001), 767–804.

1.4 On the Arithmetic Difference of Regular Cantor Sets

Sets of Small Length but Large Cardinality

Does the interval $[0, 1]$ contain more real numbers than $[0, 1/2]$? If you are seeing this question for the first time, you might answer “Yes”. Mathematicians, however, say that two sets A and B have equal cardinality (that is, the same number of elements) if there exists a one-to-one correspondence between their elements. Now, we can take any number $x \in [0, 1]$, and put it in correspondence with the number $x/2 \in [0, 1/2]$. Hence, the cardinalities of these sets are actually equal, although the lengths are different.

By a similar argument, an interval of any length $\varepsilon > 0$ has the same cardinality as $[0, 1]$. But what about even smaller lengths? A set S of real numbers has length (or measure) 0 if, for any $\varepsilon > 0$, it can be covered by a set of intervals of total length ε . For example, the set of rational numbers in $[0, 1]$ has measure 0. Indeed, for any $\varepsilon > 0$, every number m/n can be covered by an interval

$$(m/n - \varepsilon/2n2^n, m/n + \varepsilon/2n2^n)$$

of length $\varepsilon/n2^n$. In this case, the n rational numbers $1/n, 2/n, \dots, n/n$ with denominator n are covered by n such intervals with total length $n \cdot (\varepsilon/n2^n) = \varepsilon/2^n$. So, rational numbers with denominators $n = 2, 3, 4, \dots$ are covered by intervals of total length $\varepsilon/2, \varepsilon/4, \varepsilon/8, \dots$ and the total length of all intervals is bounded by $\varepsilon/2 + \varepsilon/4 + \varepsilon/8 + \dots = \varepsilon$.

It is known that one cannot create a one-to-one correspondence between the set of rational numbers and the set of real numbers. One may ask, however, if there exists a set with measure 0 which still has the same cardinality as $[0, 1]$. Again, if you are new to the subject, you might guess that there is no such set, however there is, and here is an example.

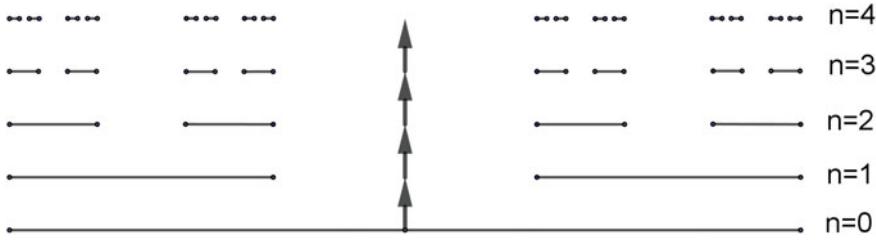


Fig. 1.4 The first four steps in the construction of the Cantor set

Cantor Sets

Take the interval $[0, 1]$, and delete the middle third $(1/3, 2/3)$, resulting in the set $[0, 1/3] \cup [2/3, 1]$. Then delete the middle third from each of these two intervals, resulting in $[0, 1/9] \cup [2/9, 1/3] \cup [2/3, 7/9] \cup [8/9, 1]$, and continue this process indefinitely, see Fig. 1.4. Let C be the set of points which is never deleted. After step 1, we have deleted an interval of length $1/3$, so the total length of the remaining part is $2/3$. At step 2, we have deleted $1/3$ of this, and the total length of the remaining part is $2/3 \cdot 2/3 = 4/9$. Similarly, the length of the part remaining after step n is $(2/3)^n$, hence the measure of C is $\lim_{n \rightarrow \infty} (2/3)^n = 0$. However, C is non-empty (in particular, one can easily check that $0 \in C$, and $1 \in C$), and in fact there is a one-to-one correspondence between C and the interval $[0, 1]$.

Indeed, take any number $x \in [0, 1]$, and assign to it the letter L or G according to whether it belongs to the subinterval $[0, 1/2]$ or $[1/2, 1]$, respectively. For example, $x = 1/3$ is assigned the letter L . Then divide the corresponding subinterval into two halves again, and assign a second letter, for example, numbers from $[0, 1/2]$ are assigned L or G if they belong to the subintervals $[0, 1/4]$ or $[1/4, 1/2]$, respectively, in particular $x = 1/3$ is assigned G this time. After repeating this infinitely, we can associate to each $x \in [0, 1]$ a unique infinite sequence of L 's and G 's. Now, in the process of constructing the set C as above, we can similarly assign to every $y \in C$ a first letter L if $y \in [0, 1/3]$ and G if $y \in [2/3, 1]$. Then, if, say, $y \in [0, 1/3]$, assign a second letter L if $y \in [0, 1/9]$ and G if $y \in [2/9, 1/3]$, and so on. In this way we associate infinite sequence of L 's and G 's to every $y \in C$. Finally, we put $x \in [0, 1]$ into correspondence with $y \in C$ if and only if x and y are associated with the same sequence of letters.

Hence, we have constructed a set C with measure (length) 0, but with the same number of elements as the full interval $[0, 1]$. Obviously, the construction above is not unique. For each $\beta \in (0, 1/2)$, we can delete the middle part $(\beta, 1 - \beta)$ of $[0, 1]$, and then repeat the process as above. Also, we can start with any interval $[a, b]$ instead of $[0, 1]$. The resulting sets all have as many elements as $[0, 1]$ but measure 0, and are examples of *Cantor sets*.

Regular Cantor Sets in Dynamical Systems

At first, Cantor sets look like exotic objects. However, they play a fundamental role in several areas of mathematics, including, for example, the study of *dynamical systems*. Given any function $f : \mathbb{R} \rightarrow \mathbb{R}$, and an initial point $x_0 \in \mathbb{R}$, we can study a sequence $x_1 = f(x_0)$, $x_2 = f(x_1) = f(f(x_0))$, $x_3 = f(x_2) = f(f(f(x_0)))$, \dots . A crucial question is whether such a sequence remains bounded, or goes to infinity. For example, take the function

$$f(x) = \begin{cases} 3x, & \text{if } x \leq 0.5, \\ 3(1-x), & \text{if } x \geq 0.5, \end{cases}$$

and $x_0 = 2$. Then $x_1 = 3(1-2) = -3$, $x_2 = 3(-3) = -9$, $x_3 = 3(-9) = -27$, \dots , $x_n = -3^n$, \dots , and the sequence decreases to $-\infty$. Actually, for any $x_0 < 0$ we have $x_n = x_0 3^n \rightarrow -\infty$, and similarly for any $x_0 > 1$, $x_1 = 3(1-x_0) < 0$, hence $x_n = x_1 3^{n-1} \rightarrow -\infty$, so the only way for the sequence to stay bounded is to remain within $[0, 1]$. Further, if $x_0 \in (1/3, 2/3)$, then $x_1 > 1$, $x_2 < 0$, $x_n = x_2 3^{n-2}$, hence we should have $x_0 \in [0, 1/3] \cup [2/3, 1]$. Similarly, for $x_0 \in (1/9, 2/9) \cup (7/9, 8/9)$, we get $x_1 \in (1/3, 2/3)$, $x_2 > 1$, again resulting in an unbounded sequence, hence $x_0 \in [0, 1/9] \cup [2/9, 1/3] \cup [2/3, 7/9] \cup [8/9, 1]$. Continuing in this way, we can conclude that the set of initial points x_0 for which the sequence stays bounded is exactly the Cantor set C defined above. Similarly, for the function $f_t(x) := f(x-t) + t$, $t \in \mathbb{R}$, the set of initial points leading to a bounded sequence is the Cantor set $C + t = \{z \mid z = y + t, y \in C\}$. More generally, any Cantor set C' arising in this way from some function g is called a *regular Cantor set*.

The Arithmetic Difference of Regular Cantor Sets

While studying dynamical systems, it is often important to find a point x_0 such that the dynamics remains bounded for two different functions at once. For example, both sequences $x_0, f_t(x_0), f_t(f_t(x_0)), \dots$, and $x_0, g(x_0), g(g(x_0)), \dots$ are simultaneously bounded if and only if $x_0 \in (C + t) \cap C'$. The set $(C + t) \cap C'$ is non-empty if and only if there exist $x \in C'$, $y \in C$ such that $t = x - y$, that is, t can be written as a difference of elements from corresponding Cantor sets.

In 1987, Palis [296] conjectured that, for generic pairs of regular Cantor sets C_1 and C_2 , their arithmetic difference $C_1 - C_2 = \{x - y \mid x \in C_1, y \in C_2\}$ is either of measure 0 or contains an interval. The word “generic” here means that the statement is true with probability 1 if C_1 and C_2 are selected “at random”. For example, the set of rational numbers in $[0, 1]$ has measure 0, therefore the set of all irrational numbers on $[0, 1]$ has measure 1. So, if a real number on $[0, 1]$ is selected at random, it would be irrational with probability 1. In other words, a generic real number is irrational.

The Proof of the Palis Conjecture

The conclusion that a set is “either of measure 0 or contains an interval” is non-trivial because there exist sets of positive measure not containing an interval. For example, we have just concluded that the set of all irrational numbers in $[0, 1]$ has measure 1, but there are no intervals $[a, b] \subset [0, 1]$ with $a < b$ such that all numbers $x \in [a, b]$ are irrational. In fact, because Cantor sets are intuitively rather “strange”, one would expect their difference also to be “strange”, and, from this point of view, the Palis conjecture even looks a bit surprising. Because the arithmetic difference of Cantor sets arises naturally in the study of dynamical systems, this was an important question which remained open for almost 15 years.

The following theorem of Moreira and Yoccoz [281] states that the Palis conjecture is true.

Theorem 1.4 *The arithmetic difference of a generic pair of regular Cantor sets either has measure zero or contains an interval.*

Actually, Moreira and Yoccoz proved a much more general result, from which they deduced the Palis conjecture as a corollary. Below we give an *informal* description of their general result.

Hausdorff Dimension and Stable Intersections

While constructing the Cantor set C in Fig. 1.4, we deleted the middle third of each interval at every step. Intuitively, if we had deleted the middle fifth of each interval instead, we would delete *less*, and the set C' of the remaining points should be in some sense “larger” than C . But in what sense? Both C and C' have the same cardinality as $[0, 1]$, but both have measure 0, so how can we say that one such set is “larger” than another one? It turns out that the correct notion to quantify how “large” Cantor sets are is the (appropriately generalized) notion of *dimension*.

We need about $1/\varepsilon$ intervals of length ε to cover the unit interval, but would require about $(1/\varepsilon)^2$ squares of side length $1/\varepsilon$ to cover the unit square, and, more generally, about $N(\varepsilon) = (1/\varepsilon)^d$ d -dimensional cubes with side length ε to cover the unit d -dimensional cube. Hence, the dimension d is equal to $\ln(N(\varepsilon))/\ln(1/\varepsilon)$. Now, we can cover the set C by $N = 2^n$ intervals of length $\varepsilon = 1/3^n$ each, see Fig. 1.4 for an illustration for $n = 0, 1, 2, 3, 4$, hence $\ln(N(\varepsilon))/\ln(1/\varepsilon) = \ln(2)/\ln(3) \approx 0.63$. This non-integer number is denoted by $d(C)$ and is called the *Hausdorff dimension* of the set C . A similar calculation shows that the Hausdorff dimension of C' is $d(C') = \ln(2)/\ln(5/2) \approx 0.76$, confirming the intuition that C' is “larger” than C .

Let C'' be the set arising from the same construction as C but where we delete a $1/3.0001$ th part of each interval instead of one third. Intuitively, C'' is much “closer” to C than C' , or, in other words, C'' lies in a “small neighbourhood” of C . We say that regular Cantor sets K_1 and K_2 have *stable intersection* if the intersection of K'_1 and K'_2 is non-empty, whenever K'_1 and K'_2 are regular Cantor sets lying in a “small

neighbourhood” of K_1 and K_2 . The general result of Moreira and Yoccoz states that, for a generic pair of regular Cantor sets K_1 and K_2 , such that $d(K_1) + d(K_2) > 1$, there exists a $t \in \mathbb{R}$ such that K_1 and $K_2 + t$ have stable intersection. Moreover, they proved that there are in fact “many” values of t satisfying this condition. They deduced Theorem 1.4 as a corollary of this result.

Reference

C.J. Moreira and J.-C. Yoccoz, Stable intersections of regular Cantor sets with large Hausdorff dimensions, *Annals of Mathematics* **154**-1, (2001), 45–96.

1.5 The Existence of Different Groups with Isomorphic Group Rings

Some Properties of Addition and Multiplication

The addition operation for integers satisfies the following useful properties:

- (i) The sum of any two integers is again an integer;
- (ii) $(a + b) + c = a + (b + c)$ for all integers a, b, c ;
- (iii) there exists a special integer 0 such that $a + 0 = 0 + a = a$ for all integers a ;
- (iv) for every integer a , there exists an integer b (namely, $b = -a$), such that $a + b = b + a = 0$.

The multiplication operation obeys similar properties for non-zero real numbers:

- (i) The product of any two non-zero real numbers is again a non-zero real number;
- (ii) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all non-zero real numbers a, b, c ;
- (iii) there exists a special non-zero real number 1 such that $a \cdot 1 = 1 \cdot a = a$ for all non-zero real a ;
- (iv) for every non-zero real number a , there exists a b (namely, $b = 1/a$), such that $a \cdot b = b \cdot a = 1$.

Note that if we consider multiplication for all real numbers, or for all integers, or for non-zero integers, then property (iv) fails.

Composition of Bijections

Now consider a slightly more complicated example. Let S be an arbitrary set. A function $f : S \rightarrow S$ is called a *bijection* if for every $y \in S$ there exists exactly one $x \in S$ such that $f(x) = y$. For example, if $S = \mathbb{R}$ is the set of all real numbers, the function $f(x) = x^3$ is a bijection, because for every $y \in \mathbb{R}$ there exists exactly one $x \in \mathbb{R}$ (namely, $x = \sqrt[3]{y}$) such that $x^3 = y$. On the other hand, the function $f(x) = 2^x$ is not a bijection because for $y = -1$ there are no real x such that $2^x = -1$. The function $f(x) = \tan(x)$ is also not a bijection, because, for example, for $y = 0$ there

are *many* x such that $\tan(x) = 0$ (for example, $x = 0$ and $x = \pi$ works), while the definition requires that there should be *exactly one* such x . If $f : S \rightarrow S$ is a bijection, then there exists a function $g : S \rightarrow S$ such that $f(x) = y \Leftrightarrow g(y) = x$. We will call g the *inverse function* to f . For example, the inverse function to $f(x) = x^3$ is $g(y) = \sqrt[3]{y}$. The inverse function to f is also denoted f^{-1} .

For any functions $f : S \rightarrow S$ and $g : S \rightarrow S$, we say that the function $h : S \rightarrow S$ is the *composition* of f and g , and write $h = g \circ f$, if $h(x) = g(f(x))$ for all $x \in S$. For example, if $S = \mathbb{R}$, $f(x) = x^3$, and $g(x) = x^5$, then $h(x) = (x^3)^5 = x^{15}$. Then the following properties hold.

- (i) The composition of any two bijections is again a bijection;
- (ii) $(h \circ g) \circ f = h \circ (g \circ f)$ for all bijections f, g, h ;
- (iii) there exists a bijection e (namely, $e(x) = x$, $\forall x \in S$) such that $f \circ e = e \circ f = f$ for all f ;
- (iv) for every bijection f , there exists a bijection g (namely, $g = f^{-1}$), such that $g \circ f = f \circ g = e$.

Groups

We could continue with hundreds of examples like this, from all areas of mathematics. We have some “objects” (integers, non-zero real numbers, bijections, etc.), and some operation we can perform with them (addition, multiplication, composition, etc.), such that properties (i)–(iv) hold. It is convenient to have one general definition which generalizes all such examples, and this is the definition of a *group*. A group is any set G and operation \cdot which assigns to any two elements $a, b \in G$ a third one, denoted $a \cdot b$, or just ab , such that the following properties hold.

- (i) $a \cdot b \in G$ for all $a, b \in G$;
- (ii) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all $a, b, c \in G$;
- (iii) there exists an $e \in G$ such that $a \cdot e = e \cdot a = a$ for all $a \in G$;
- (iv) for every $a \in G$, there exists an element $b \in G$ (also denoted a^{-1}) such that $a \cdot b = b \cdot a = e$.

In particular, the set of all integers with the addition operation forms a group, which we will denote by \mathbb{Z} ; the set of all non-zero real numbers, denoted $\mathbb{R}/0$, forms a group with the multiplication operation, while set of all bijections $f : S \rightarrow S$ forms a group with the composition operation. Hence, if we prove any theorem about groups, it will automatically hold in all these examples, and in hundreds of others.

A group G is called *finite* if G is a finite set, and *infinite* otherwise. The groups \mathbb{Z} and $\mathbb{R}/0$ are infinite, while the group of all bijections $f : S \rightarrow S$ is infinite if the set S is infinite. However, if S is a finite set $\{1, 2, \dots, n\}$, there is a finite number³ of bijections, which forms a finite group.

³Namely, $n!$ bijections. Indeed, $f(1)$ can take n possible values, $f(2)$ $n - 1$ values (all except $f(1)$), $f(3)$ $n - 2$ values, and so on, so the total number of possible bijections f is $n \cdot (n - 1) \cdot (n - 2) \cdots 1 = n!$

Isomorphic Groups

Let e and o be the sets of all even and all odd integers, respectively. As you know, the sum of two even integers is always even, so let us write $e + e = e$. For similar reasons, $e + o = o$, $o + e = o$, and $o + o = e$. It is straightforward to verify that properties (i)–(iv) hold for the set $G_{\text{parity}} = \{e, o\}$ with operation $+$, hence this is an example of a group with just two elements.

As another example, consider the group of all bijections $f : S \rightarrow S$ in the case when the set S consists of just two elements, say, x and y . In this case, there are just two possible bijections: one (call it e) sends x to x and y to y , and another one (call it g), switches the elements: $g(x) = y$ and $g(y) = x$. Then all possible compositions are $e \circ e = g \circ g = e$ and $g \circ e = e \circ g = g$, and the set $G_{\text{comp}} = \{e, g\}$ with the composition operation \circ forms a group.

We can see that if, in G_{parity} , we rename element o as g , and the operation $+$ as the composition operation \circ , we get exactly G_{comp} . In fact, G_{parity} and G_{comp} is the same group, just expressed in two different ways. In general, we say that two groups G and H are *isomorphic* if there exists a one-to-one correspondence between their elements which preserves the group operations.

Rings

Groups are convenient if we study one operation at a time. However, what if we want to study *two* operations? For example, for the set of integers, we may study addition and multiplication, which satisfy the following properties:

- properties (i)–(iv) of $+$, listed above;
- (v) $a + b = b + a$ for all a, b ;
- (vi) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all a, b, c ;
- (vii) there exists a special element 1 such that $a \cdot 1 = 1 \cdot a = a$ for all a ;
- (viii) $a \cdot (b + c) = a \cdot b + a \cdot c$ for all a, b, c ;
- (ix) $(b + c) \cdot a = b \cdot a + c \cdot a$ for all a, b, c .

It is easy to check that addition and multiplication of real numbers, or, for example, of all functions $f : \mathbb{R} \rightarrow \mathbb{R}$, satisfy the same properties (i)–(ix). In general, any set R together with two operations $+$ and \cdot satisfying properties (i)–(ix) is called a *ring*.

Integral Group Rings and Higman's Question

For any (finite) group G , an *integral group ring* is the set of all possible sums $\sum_{g \in G} a_g g$, where a_g are integers, addition is defined as

$$\sum_{g \in G} a_g g + \sum_{g \in G} b_g g = \sum_{g \in G} (a_g + b_g) g,$$

and multiplication is given by

$$(\sum_{g \in G} a_g g) \cdot (\sum_{h \in G} b_h h) = \sum_{g \in G} \sum_{h \in G} a_g b_h g h.$$

For example, the integral group ring of $G_{\text{comp}} = \{e, g\}$ contains elements $x = e + 3g$ and $y = 2e - 5g$, their sum is $x + y = 3e - 2g$, and their product is

$$xy = (e + 3g)(2e - 5g) = 2ee - 5eg + 6ge - 15gg = 2e - 5g + 6g - 15e = -13e + g.$$

Similarly to groups, two rings R and Q are called *isomorphic* if there exists a one-to-one correspondence between their elements which preserves both operations. Obviously, if two groups G and H are isomorphic, then so are their integral group rings. In 1940, Higman [199] asked if the converse is true, that is, whether different finite groups have different integral group rings.

Why This Question Is Important

To explain the importance of this question, let us first consider an example from geometry. Assume that we have a triangle ABC in the plane, and a device for measuring the distance between any two points. Then we can measure side lengths $|AB| = a$, $|BC| = b$, $|CA| = c$, and these three numbers completely describe the triangle, because two triangles with the same side lengths are equal. This implies that we can answer all questions about triangles using just three numbers a, b, c and nothing else. For example, when asked if $\angle BAC = 90^\circ$, we can use Pythagoras' Theorem and check if $a^2 = b^2 + c^2$.

Now, if we are given a quadrangle $ABCD$, can we describe it similarly with just four numbers, representing its side lengths? For example, if $|AB| = |BC| = |CD| = |DA| = 1$, is $\angle BAC = 90^\circ$? Unfortunately, this question cannot be answered based on side lengths only. It may be that $ABCD$ is a square, and then the answer is “yes”, but it may be that it is a rhombus with $\angle BAC \neq 90^\circ$. The problem is that, in contrast to the triangle case, there are many different quadrangles with the same side lengths, see Fig. 1.5.

A positive answer to Higman's question would imply that any finite group could be completely described by its integral group ring in a similar way as three side lengths describe a triangle. Hence, any question about any finite group could be reduced to ring theory, in the same way as we have reduced any geometric question about triangles to algebra. This would be extremely useful, because lots of tools and results from ring theory could then be applied to study finite groups via the corresponding integral group rings.

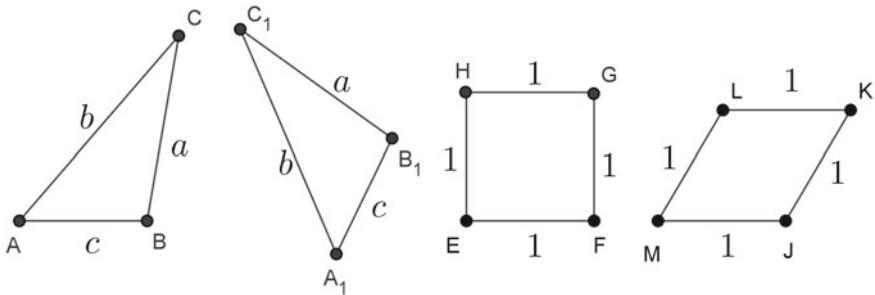


Fig. 1.5 Triangles can be described by side lengths, while quadrilaterals cannot

A Negative Answer to Higman's Question

Higman's question turned out to be very difficult and was open for more than 60 years. A positive answer has been obtained for many special classes of groups, in the form “If finite groups G and H , which have some given special properties, are different, then so are their integral group rings”, but all attempts to extend such results to the general case remained elusive. And now we know why: the following theorem of Hertweck [198] states that in the general case the answer to this question is negative.

Theorem 1.5 *There exist two finite non-isomorphic groups G and H such that their integral group rings are isomorphic.*

Theorem 1.5 shows that, by describing a group G using its integral group ring, we may lose some important information about the group in the same way as we lose information about a quadrangle by describing it using side lengths only.

Hertweck proved Theorem 1.5 by constructing such groups G and H , with $2^{21} \cdot 97^{28}$ elements each. The groups G and H are different, but in some magical way their corresponding integral group rings (that is, all the sums in the form $\sum_{g \in G} a_g g$ and $\sum_{h \in H} a_h h$, having $2^{21} \cdot 97^{28}$ terms each) completely coincide! These two huge rings turned out to be just the same ring, described in two different ways, in the same way as G_{parity} and G_{comp} turned out to be the same group. As you can guess from the number of elements, the construction is highly non-trivial. To the best of our knowledge, no smaller examples are currently known.

Reference

M. Hertweck, A counterexample to the isomorphism problem for integral group rings, *Annals of Mathematics* **154**-1, (2001), 115–138.

1.6 Short Representations of Elements of Finite Simple Groups

Groups and Subgroups

A *group* is a set G together with an operation \cdot such that (i) $a \cdot b \in G$ for all $a, b \in G$; (ii) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all $a, b, c \in G$; (iii) there exists an $e \in G$ (called the identity element of G) such that $a \cdot e = e \cdot a = a$ for all $a \in G$; and (iv) for every $a \in G$, there exists an element $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e$. For example, the set of integers \mathbb{Z} with the addition operation “ $+$ ” forms a group, and so does the set of all bijections $f : S \rightarrow S$ with the composition operation \circ , see Sect. 1.5 for details and more examples.

A group G is called *abelian* if $a \cdot b = b \cdot a$ for all $a, b \in G$. For example, the group \mathbb{Z} of integers is abelian, because $a + b = b + a$ for any integers a, b . However, for compositions, it can be that $f \circ g \neq g \circ f$. For example, if $f(x) = x + 1$ and $g(x) = x^3$, then $f \circ g = x^3 + 1 \neq (x + 1)^3 = g \circ f$. Hence, the corresponding group is not abelian.

An important way to understand the structure of a group is to study its *subgroups*. A subset H of a group G is called a *subgroup* if

- (i) $a \cdot b \in H$ for all $a, b \in H$;
- (ii) $e \in H$, where e is the identity element of G ;
- (iii) $a^{-1} \in H$ for every $a \in H$.

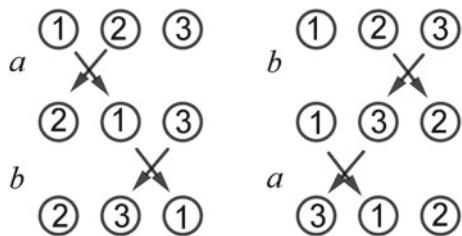
For example, the sum of any two even integers is again even, 0 is even, and $-a$ is even for every even a , hence the set of all even integers is a subgroup of \mathbb{Z} .

Symmetric Groups

If S is a finite set $\{1, 2, \dots, n\}$, any function $f : S \rightarrow S$ can be described by just listing its values, and we will write $f : (1, 2, \dots, n) \rightarrow (f(1), f(2), \dots, f(n))$, or just $f = (f(1), f(2), \dots, f(n))$. In this case, f is a bijection if and only if the set $(f(1), f(2), \dots, f(n))$ lists all the numbers from 1 to n , but possibly in a different order. In other words, f just *permutes* the numbers $1, 2, \dots, n$. For example, if $n = 3$ and f is such that $f(1) = 2$, $f(2) = 3$, $f(3) = 2$, we write $f = (2, 3, 2)$, and this f is not a bijection. In contrast, if $f(1) = 2$, $f(2) = 3$, $f(3) = 1$, we write $f = (2, 3, 1)$, and this is a bijection/permuation. In this case, the group of all bijections $f : S \rightarrow S$ is called a *symmetric group*, and is usually denoted S_n . It is a finite group with $n!$ elements. For example, the group S_3 contains $3! = 1 \cdot 2 \cdot 3 = 6$ elements: $a : (1, 2, 3) \rightarrow (2, 1, 3)$, $b : (1, 2, 3) \rightarrow (1, 3, 2)$, $c : (1, 2, 3) \rightarrow (2, 3, 1)$, $d : (1, 2, 3) \rightarrow (3, 1, 2)$, $e : (1, 2, 3) \rightarrow (1, 2, 3)$, and $f : (1, 2, 3) \rightarrow (3, 2, 1)$.

A permutation $f \in S_n$ is called *even* if the number of pairs (i, j) such that $i < j$ but $f(i) > f(j)$ is even. For example, if $f(k) = k$, $k = 1, 2, \dots, n$, then the number of such pairs is 0, hence f is even. If $f(1) = 2$, $f(2) = 1$, $f(k) = k$, $k = 3, \dots, n$,

Fig. 1.6 Illustration of $a \cdot b = c$ and $b \cdot a = d$ in S_3



then there is exactly one such pair, namely $(1, 2)$, hence f is not even. One can check that the set of all even permutations is a subgroup of S_n , which is usually denoted A_n . For example, A_3 contains 3 elements: c , d , and e .

Normal Subsets and Simple Groups

A subset T of a group G is called *normal* if $g \cdot a \cdot g^{-1} \in T$ for any $a \in T$ and $g \in G$. If G is abelian, then $g \cdot a \cdot g^{-1} = g \cdot g^{-1} \cdot a = e \cdot a = a \in T$, hence every subset T is normal. If G is not abelian, then some subsets are normal, while some are not. For example, in the group S_3 , let us prove that $a \cdot b = c \neq d = b \cdot a$. Indeed, the permutation $a : (1, 2, 3) \rightarrow (2, 1, 3)$ permutes the first two elements, while the permutation $b : (1, 2, 3) \rightarrow (1, 3, 2)$ permutes the last two. If we start with $(1, 2, 3)$, permute the first two elements, and then permute the last two, we get $(2, 3, 1)$, see Fig. 1.6. However, the permutation $(1, 2, 3) \rightarrow (2, 3, 1)$ is denoted by c , hence $a \cdot b = c$. The fact that $b \cdot a = d$ can be proved in a similar way, see Fig. 1.6. Hence, $a \cdot b \neq b \cdot a$, thus $a \cdot b \cdot a^{-1} \neq b \cdot a \cdot a^{-1} = b$, and, by definition, the set $S = \{b\}$ is not normal. However, one can check that sets $\{e\}$, $\{c, d\}$, and $\{a, b, f\}$ are all normal, as well as their unions. Actually, the set $\{e\}$ consisting of only the identity element, as well as G itself, are always normal subsets of G .

A subgroup H of group G is called *trivial* if either $H = G$, or $H = \{e\}$, and *non-trivial* otherwise. A subgroup H of a group G is called *normal* if H is at the same time a subgroup and a normal subset of G . For example, $\{e\}$, $A_3 = \{c, d, e\}$, and $S_3 = \{a, b, c, d, e, f\}$ are all normal subgroups of S_3 . In general, one can check that, for every $n \geq 5$, A_n is a normal subgroup of S_n , but all non-trivial subgroups of A_n are *not* normal.

A group G is called *simple* if it does not have any non-trivial normal subgroups. For example, the groups \mathbb{Z} and S_n ($n \geq 3$) are not simple (with non-trivial normal subgroups being the even integers and A_n , respectively), while A_n , $n \geq 5$, is an example of a simple group.

How to Write Down Group Elements

The group S_3 has only 6 elements, and we introduce a separate notation a, b, c, d, e, f for each of them. However, the group S_4 of permutation of 4 numbers contains 24 elements (there are 4 ways to choose the first number, 3 ways for the second one, and 2 ways for the third, $4 \cdot 3 \cdot 2 = 24$), the group S_5 has 120 elements ($5 \cdot 4 \cdot 3 \cdot 2 = 120$), and so on, so if we wanted to continue to give every element a symbol, any alphabet we choose would quickly be exhausted. Hence, another way of representing elements of a group needs to be developed.

In S_3 , we can take a normal set $S = \{a, b, f\}$, and then express all elements of S_3 as a product of at most two elements of S

$$a = a, \quad b = b, \quad c = ab, \quad d = af, \quad e = aa, \quad f = f. \quad (1.7)$$

In this way, all elements of G_3 can be expressed using just 3 letters. Can we use fewer? Taking $S = \{c, d\}$, we can write $e = cd$, and... that's all: $dc = e$, $cc = d$, $dd = c$, $de = ed = d$, $ce = ec = c$, so there is no way to represent, for example, a . This is because $S = \{c, d\}$ is contained in a subgroup $A_3 = \{c, d, e\}$, and we cannot represent any element outside A_3 .

In general, if a set S is contained in some proper subgroup H of G , then, by writing different products of elements from S , we cannot go outside of H , and therefore cannot express all elements of G .

Representing Elements of a Simple Group

By definition, a simple group does not have any normal subgroups except for $\{e\}$ and itself. Therefore, if S is any normal subset of a simple group G , it cannot be a subgroup, the situation described above cannot happen, and therefore any element of $g \in G$ can be represented as

$$g = s_1 s_2 \dots s_m \quad (1.8)$$

for some natural m and $s_1, \dots, s_m \in S$. However, if representation (1.8) is too long, it is not very useful. The main question is therefore *how many* elements of S is needed to represent every $g \in G$?

Let $|S|$ and $|G|$ be the number of elements in S and G , respectively. For any m , there are exactly $|S|^m$ ways to write a product in the form $s_1 s_2 \dots s_m$ (there are $|S|$ ways to choose $s_1 \in S$, $|S|$ ways to choose $s_2 \in S$, and so on). Hence, every element of G can be represented in the form (1.8) only if $|S|^m \geq |G|$, which implies

$$m \geq \ln |G| / \ln |S|. \quad (1.9)$$

This estimate, however, is too optimistic: $m = \ln |G| / \ln |S|$ can only hold if all products $s_1 s_2 \dots s_m$ represent different elements of G . However, what can prevent

these products from covering some elements of G multiple times, while leaving some others uncovered? Because of this effect, we actually need m to be larger than $\ln |G| / \ln |S|$ to guarantee that all elements of G are covered. But how much larger? The following theorem of Liebeck and Shalev [242] states that $m \geq c \ln |G| / \ln |S|$ for some constant c suffices.

Theorem 1.6 *There exists a constant c such that if G is a finite simple group and $S \neq \{e\}$ is a normal subset of G , then, for any $m \geq c \ln |G| / \ln |S|$, any element of G can be expressed as a product of m elements of S .*

In general, understanding the structure of finite simple groups plays a very important role in group theory. There is a theorem stating that any finite group can be decomposed into uniquely determined simple groups, hence simple groups are the building blocks for all finite groups in a similar way as prime numbers are the building blocks for all integers.

It follows from (1.9) that the inequality $m \geq c \ln |G| / \ln |S|$ in Theorem 1.6 is the best possible up to a constant factor. The fact that any element of G can be written as a product of a small number of elements from any normal subset S is beautiful by itself, but it also turned out to be extremely useful. In fact, Liebeck and Shalev derived ten interesting and non-trivial corollaries from it in their paper, not counting further applications. One such corollary is stated below.

For any finite simple group G and normal subset $S \neq \{e\}$ of it, and any two elements $a \in G$ and $b \in G$, define the “distance” $d_{G,S}(a, b)$ to be the minimal integer k such that $b = as_1s_2 \dots s_k$ for some $s_i \in S$, $i = 1, 2, \dots, k$. That is, $d_{G,S}(a, b)$ is the minimal number of multiplications by elements of S required to “transform” a into b . Let the diameter $\text{diam}(G, S)$ be the maximal possible value of $d_{G,S}(a, b)$ for $a \in G$ and $b \in G$. Then Theorem 1.6 implies that $\text{diam}(G, S) \leq c \ln |G| / \ln |S|$.

Reference

M. Liebeck and A. Shalev, Diameters of finite simple groups: sharp bounds and applications, *Annals of Mathematics* **154**-2, (2001), 383–406.

1.7 The Existence of a Field with u -Invariant 9

The Field of Rational Numbers

Historically, people first studied the set \mathbb{N} of natural numbers $0, 1, 2, \dots$, and basic operations over them: addition, multiplication, and their inverses, subtraction and division. The results of the latter two operations, however, are not always natural numbers, therefore negative numbers and fractions were introduced, resulting in the set \mathbb{Q} of rational numbers. Within \mathbb{Q} , all four basic operations $+, -, \cdot$, and $/$, are well-defined, except for division by 0. A set on which these operations can be defined, except for division by a special element called 0, and for which some standard basic properties hold, is called a *field*.

More formally, a field is a set F together with operations $+$ and \cdot such that

- (i) $a + b \in F$ and $a \cdot b \in F$ for every $a, b \in F$;
- (ii) $(a + b) + c = a + (b + c)$ and $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all $a, b, c \in F$;
- (iii) $a + b = b + a$ and $a \cdot b = b \cdot a$ for all $a, b \in F$;
- (iv) there exist two different elements of F , usually denoted 0 and 1, such that
 $a + 0 = a$ and $a \cdot 1 = a$ for all $a \in F$;
- (v) for every $a \in F$, there exists an element in F , denoted $-a$, such that $a + (-a) = 0$;
- (vi) for every $a \neq 0$ in F , there exists an element in F , denoted a^{-1} or $1/a$, such that $a \cdot a^{-1} = 1$;
- (vii) $a \cdot (b + c) = a \cdot b + a \cdot c$ for all $a, b, c \in F$.

In any field F , subtraction is defined as $a - b = a + (-b)$ and division is defined as $a/b = a \cdot (1/b)$.

More Examples of Infinite and Finite Fields

In the field \mathbb{Q} of rational numbers, however, one cannot, for example, extract the square root of 2. The latter can be written in the form $\sqrt{2} = 1.41421\dots$ of an infinite decimal, and the set of all infinite decimals forms the field \mathbb{R} of real numbers. Further, even within the field \mathbb{R} one cannot extract a square root of -1 . The latter is called an imaginary number, usually denoted $i = \sqrt{-1}$, and the set of all numbers of the form $a + ib$, $a, b \in \mathbb{R}$, forms the set \mathbb{C} of complex numbers. Within \mathbb{C} , we can naturally define the basic operations as

$$(a + ib) + (c + id) = (a + c) + i(b + d), \quad (a + ib) - (c + id) = (a - c) + i(b - d), \quad (1.10)$$

$$(a + ib)(c + id) = ac + iad + ibc + i^2bd = (ac - bd) + i(ad + bc), \quad (1.11)$$

$$\frac{a + ib}{c + id} = \frac{(a + ib)(c - id)}{(c + id)(c - id)} = \frac{ac - iad + ibc - i^2bd}{c^2 - i^2d^2} = \frac{ac + bd}{c^2 + d^2} + i \frac{bc - ad}{c^2 + d^2}, \quad (1.12)$$

and one may verify that properties (i)–(vii) hold, so that \mathbb{C} is actually a field.

In the opposite direction, there exist fields smaller than \mathbb{Q} , in fact even with a finite number of elements. First, think about time measured in hours. If we add 3 h to 11.00 pm, we get 2.00 am, so $11 + 3 = 2$, not 14. Similarly, 5 h before 3 am is 10 pm, so $3 - 5 = 10$, not -2 . In general, if the result of any operation is outside the range from 1 to 12 (or, equivalently, from 0 to 11 if we call midnight 0.00), we add or subtract 12 to return within the range. In this way we can always define addition, subtraction, and multiplication, for example $3 \cdot 7 = 9$ (after midnight, wait for 7 h three times, and the time will be 9.00 pm), see Fig. 1.7. Similarly, $3 \cdot 4 = 0$, $6 \cdot 4 = 0$, etc. However, from the last two expression we get $3 = 0/4$ and $6 = 0/4$, hence $3 = 6$, a contradiction, so in fact division is not well defined. This is because property (vi) fails for, say, $a = 2$: for any integer x , $a \cdot x$ is even, hence there is no x with $a \cdot x = 1$.

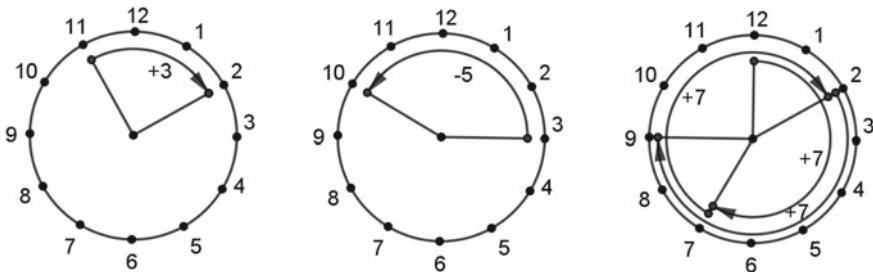


Fig. 1.7 Illustration of clock arithmetic

This problem arises because 12 can be non-trivially factorized, e.g. $12 = 4 \cdot 3$. However, if we build a similar arithmetic based on any prime number p , we really get a finite field with p elements, denoted F_p . Say, for $p = 3$, we have three elements, 0, 1, 2, and operations such that, say, $2 + 2 = 1$, so $1 - 2 = 2$, and similarly $2 \cdot 2 = 1$, hence $1/2 = 2$, etc, and one may verify that all properties (i)–(vii) hold.

Solving Quadratic Equations in Fields

Maybe the most natural thing to do in a field F is to solve equations, and the simplest equations are linear, that is, of the form $\sum_{i=1}^n a_i x_i = 0$, where x_i are variables and $a_i \in F$ are non-zero coefficients. However, linear equations are trivial, because we can always assign x_2, \dots, x_n to be arbitrary and take $x_1 = -\frac{a_2 x_2 + \dots + a_n x_n}{a_1}$. The next case is quadratic equations of the form

$$\sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j = 0. \quad (1.13)$$

In this case, we always have the trivial solution $x_1 = x_2 = \dots = x_n = 0$, but can we always find at least one other solution, called non-trivial? In general, we cannot. For example, a quadratic equation in any number of variables $x_1^2 + x_2^2 + \dots + x_n^2 = 0$ over the field of rationals or reals has no non-trivial solutions. Indeed, all x_i^2 , $i = 1, 2, \dots, n$, are non-negative, and a sum of non-negative numbers can be zero only if all those numbers are equal to 0.

The situation is completely different with complex numbers. The equation in one variable $ax^2 = 0$, $a \neq 0$, indeed has no non-trivial solution, but the equation $x^2 + y^2 = 0$ can be solved by taking $y = 1$, and finding x from equation $x^2 + 1 = 0$, which has solutions $x = i$ and $x = -i$. Similarly, any equation of the form $ax^2 + bxy + cy^2 = 0$ (where a, b, c are some constants) has a non-trivial solution with $y = 1$ and x being a root of the corresponding quadratic equation $ax^2 + bx + c = 0$, which can be found from the usual formula

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

because in the field of complex numbers the square root always exists.

The u -Invariant of a Field

In general, a number $u(F)$ is called a *u -invariant* of a field F if

- (a) Equation (1.13) in $u(F)$ variables may have no non-trivial solutions, but
- (b) Equation (1.13) in $u(F) + 1$ variables *always* has a non-trivial solution.

We have just proved that, for the field \mathbb{C} of complex numbers, Eq. (1.13) in one variable may have no non-trivial solutions, but Eq. (1.13) in two variables always has a non-trivial solution. By definition, this means that $u(\mathbb{C}) = 1$. More generally, a field F is called *quadratically closed* if for any $x \in F$ there exists a $y \in F$ such that $y^2 = x$. Then the same proof shows that $u(F) = 1$ for any quadratically closed field F .

For any prime p , there are equations of the form (1.13) with $n = 2$ variables in the finite field F_p that have no non-trivial solutions. For example, in F_3 one such equation is $x^2 + y^2 = 0$. However, the equation $x^2 + y^2 + z^2 = 0$ has a non-trivial solution $x = y = z = 1$ in F_3 , and this remains true for any equations of the form (1.13) with $n = 3$ variables in any field F_p . Hence, $u(F_p) = 2$.

It follows from (1.10) to (1.12) that if $x = a + ib$ and $y = c + id$ are complex numbers such that a, b, c, d are all rational, then the numbers $x + y, x - y, xy$, and x/y (if $y \neq 0$) can also be represented as $e + if$ for some rational e, f . Hence, the set of all such numbers also forms a field, usually denoted $\mathbb{Q}[i]$. It is known that a non-trivial solution to (1.13) over $\mathbb{Q}[i]$ always exists for $n = 5$, but may not exist for $n = 4$, so that $u(\mathbb{Q}[i]) = 4$.

A *monomial* in n complex variables z_1, \dots, z_n is a function of the form $f(z_1, \dots, z_n) = az_1^{k_1}z_2^{k_2}\dots z_n^{k_n}$, where $a \in \mathbb{C}, k_1, \dots, k_n \in \mathbb{N}$. A *polynomial* is a sum of any number of monomials, and a *rational function* is a ratio of two polynomials. An example of a rational function is $\frac{z_1^2+1}{2z_2^2+z_3}$. It is easy to see that the sum, difference, product and ratio of two rational functions is again a rational function, hence they form a field. It turns out that the u -invariant of such a field is 2^n .

What Numbers Can Be u -Invariants of a Field?

In all the examples above, the u -invariant of a field is always a power of 2. In 1953, Kaplansky [215] conjectured that in fact the u -invariant of any field is always a power of 2. This conjecture was widely believed, proved for many special cases, but then disproved by A. Merkurev [273] in 1989, who constructed a field with u -invariant 6. Later, he proved [274] that for any even number $2k$ there exists a field with u -invariant $2k$. On the other hand, it is known [235, Proposition 4.8] that a u -invariant

can never be 3, 5, or 7, so people started to think that it can never be an odd number greater than 1. So, the following theorem of Oleg Izhboldin [209] is to some extent a surprise.

Theorem 1.7 *There exists a field F with u -invariant 9.*

In other words, Izhboldin constructed a field F in which all equations of the form (1.13) in $n = 10$ variables always have a non-trivial solution, but this is not true for $n = 9$. After this work, the basic question “What natural numbers can be u -invariants of a field?” looks even more mysterious, and we presently do not even have a plausible conjecture about it.

Reference

O. Izhboldin, Fields of u -invariant 9, *Annals of Mathematics* **154**-3, (2001), 529–587.

Chapter 2

Theorems of 2002



2.1 Counting Rational Functions with Given Critical Points

Rational Functions

A *real rational function* of degree d is a function of the form $f(x) = P(x)/Q(x)$, where $P(x)$ and $Q(x)$ are polynomials of degree d with real coefficients. Like usual fractions, rational functions can sometimes be simplified by cancellation, for example,

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

For simplicity, we will further assume that rational functions are given in a simplified form, that is, $P(x)$ and $Q(x)$ have no common factors that can be cancelled out.

In the simplest case $d = 1$ any rational function has the form

$$l(x) = \frac{ax + b}{cx + e}, \quad (2.1)$$

where a, b, c, e are any numbers such that $ae - bc \neq 0$. Indeed, linear polynomials $ax + b$ and $cx + e$ can have a common factor only if $ax + b = k(cx + e)$, but in this case $a = kc$, $b = ke$, hence $ae - bc = (kc)e - (ke)c = 0$.

Critical Points of Rational Functions

A *critical point* of a differentiable function is a point where its derivative is equal to 0. The set of critical points is important because it contains all local minima and maxima of the function. For a rational function, the derivative

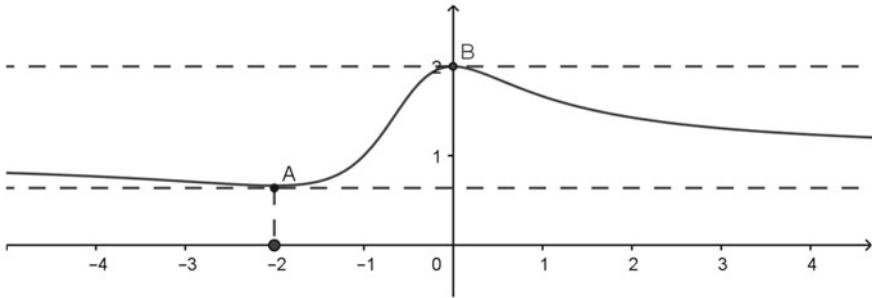


Fig. 2.1 Critical points of $g(x)$ given by (2.2)

$$f'(x) = \frac{P'(x)Q(x) - P(x)Q'(x)}{Q^2(x)}$$

is equal to 0 if $P'(x)Q(x) - P(x)Q'(x) = 0$. For $d = 1$, the function (2.1) has no critical points, because $l'(x) = 0$ would imply $(ax + b)c - a(cx + e) = 0$, or $bc - ae = 0$, a contradiction, hence the first interesting case is $d = 2$. For example, the critical points of the rational function

$$g(x) = \frac{x^2 + 2x + 2}{x^2 + x + 1} \quad (2.2)$$

satisfy the equation $(2x + 2)(x^2 + x + 1) - (x^2 + 2x + 2)(2x + 1) = 0$, and are $x_1 = -2$ and $x_2 = 0$, see Fig. 2.1.

Can We Recover a Rational Function from Its Critical Points?

In applications, we can sometimes observe the behaviour of some unknown function f describing a real-world phenomenon, compute all its local minima and maxima, and then try to find a rational function of a given degree which describes the phenomenon as accurately as possible, and in particular has the same critical points as computed. So, the crucial question is: what information about a rational function can be recovered given only its degree and its set of critical points? For example, given that $d = 2$ and the critical points are -2 and 0 , can we conclude that the function is necessarily given by (2.2)? Actually, no, there are other functions with the same set of critical points.

For example, let $h(x)$ be the composition of $g(x) = P(x)/Q(x)$ and any rational function l of degree 1 given by (2.1), that is,

$$h(x) = l(g(x)) = \left[a \frac{P(x)}{Q(x)} + b \right] / \left[c \frac{P(x)}{Q(x)} + e \right] = \frac{aP(x) + bQ(x)}{cP(x) + eQ(x)}. \quad (2.3)$$

For example, with $g(x)$ given by (2.2),

$$h(x) = l(g(x)) = \frac{(a+b)x^2 + (2a+b)x + (2a+b)}{(c+e)x^2 + (2c+e)x + (2c+e)}. \quad (2.4)$$

Because the derivative of l is never zero, the derivative of $h(x)$, given by the formula $h'(x) = l'(g(x))g'(x)$, is equal to 0 if and only if $g'(x) = 0$, hence the sets of critical points of h and g coincide. If $h(x) = l(g(x))$ as above, we will say that the functions g and h are *equivalent*.

Equivalence Classes of Rational Functions

Now, the question is, are all rational functions of degree 2 with critical points -2 and 0 given by (2.4), or are there other functions with the same set of critical points? If g_2 were such a function, then we would immediately get the whole family of new solutions in the form $h(x) = l(g_2(x))$ with l given by (2.1), that is, a family of solutions equivalent to g_2 . Assume that we continue in this way and discover that there are no more such functions after step k . Then we could describe all rational functions with a given set of critical points as “functions g, g_2, g_3, \dots, g_k and all functions equivalent to them”. The families of equivalent functions are called *equivalence classes*, and we need to list just one representative from each class to completely describe it.

However, how can we know that the number of equivalence classes is finite, that is, that we will finally find all solutions in this way after some step k ? Even if it is finite, can we estimate how many steps we may need? In 1991, Lisa Goldberg [163] proved a useful upper bound

$$k(d) \leq u_d := \frac{1}{d} \frac{(2d-2)!}{(d-1)!(d-1)!},$$

where $k(d)$ is the number of equivalence classes of rational functions of degree d . For $d = 2$, this gives $k \leq \frac{1}{2} \frac{1 \cdot 2}{1 \cdot 1} = 1$, so in fact there is only one equivalence class, and the set of all rational functions of degree 2 with critical points -2 and 0 is indeed given by (2.4). However, the number u_d grows quickly with d (for example, $u_{10} = 4862$, $u_{20} = 1767263190$), so the question arises if the bound $k(d) \leq u_d$ is the best possible, or maybe we can improve it? A theorem of Alex Eremenko and Andrei Gabrielov [141] answers this question.

Theorem 2.1 *Given $2d - 2$ distinct real points, there exists at least u_d equivalence classes of real rational functions of degree d with these critical points.*

In combination with Goldberg's result, Theorem 2.1 states that, under the conditions of the theorem, $k(d) = u_d$, so we now know *exactly* how many equivalence classes there are.

Applications to the Complex Case

Theorem 2.1 has important applications for the more general class of rational functions defined over the set \mathbb{C} of complex numbers (that is, numbers of the form $z = a + bi$, where a, b are real numbers, and $i = \sqrt{-1}$). A complex polynomial of degree d is an expression of the form $P(z) = a_0 + a_1 z + \cdots + a_d z^d$, where z is a complex variable, and $a_0, a_1, \dots, a_d \neq 0$ are complex coefficients. A complex *rational function* of degree d is a ratio of two such polynomials, $f(z) = P(z)/Q(z)$. A complex number z_0 is a critical point of f if $f'(z_0) = 0$. Again, the question is, what can we say about the function f knowing only information about its critical points? For example, what if all critical points $z_k = a_k + b_k i$ of f are real numbers, that is, $b_k = 0$? Obviously, f can be a real rational function, such as $g(x)$ above. Also, it can be equivalent to a real rational function, because critical points of equivalent rational functions coincide. Applying Theorem 2.1, Eremenko and Gabrielov derived the following result: *If all critical points of a rational function f are real, then f is equivalent to a real rational function.*

With $f(z) = P(z)/Q(z)$, the last statement can be equivalently reformulated as: *If for polynomials $P(z), Q(z)$ all the solutions of the equation $P(z)Q'(z) - P'(z)Q(z) = 0$ are real, then $P(z)$ and $Q(z)$ can be made real by a linear transformation with constant coefficients, that is, $aP(x) + bQ(x)$ and $cP(x) + eQ(x)$ are real polynomials for some $a, b, c, e \in \mathbb{C}$ such that $ae - bc \neq 0$.* This is a special case of a well-known conjecture of B. and M. Shapiro, stating that a similar result holds for any number of polynomials. This conjecture is important in the field of mathematics called Real Enumerative Geometry.

A Simplified Proof

The original proof of Theorem 2.1 in [141] is quite complicated, and uses some advanced mathematics. In 2005, Eremenko and Gabrielov [142] found a simpler proof. An Internet version is available at <http://www.math.purdue.edu/~eremenko/dvi/newshapiro5.pdf>. If you want to understand the proof of the theorem, please follow the link and read it!

Reference

- A. Eremenko and A. Gabrielov, Rational functions with real critical points and the B. and M. Shapiro conjecture in real enumerative geometry, *Annals of Mathematics* **155**-1, (2002), 105–129.

2.2 Representing Braids as Matrices

Braids and Their Products

Let X_1 be a set of m points in the coordinate plane with coordinates $(x, 1), (x, 2), \dots, (x, m)$, respectively, and X_2 be a parallel set of points with coordinates $(y, 1), (y, 2), \dots, (y, m)$ for some $y > x$. Assume that we connect points from X_1 to some points from X_2 using m (not necessarily straight) strands, going from left to right, such that no strands have a common end. If two strands intersect, one should be put above the other, and this then cannot be changed without destroying the strands. This whole configuration is called a *braid*. The exact form of the strands, as well as the exact values of x and y , are not important: if one configuration can be transformed to another without destroying the strands, it is considered to be the same braid, see Fig. 2.2. What is important, is how the strands are intertwined.

Let X be a braid as above, and let Y be another braid, going from points in X_2 to a set of points X_3 with coordinates $(z, 1), (z, 2), \dots, (z, m)$ for some $z > y > x$. Then X and Y forms another braid Z , consisting of m strands going from points in X_1 to points in X_3 . Then we call Z the product of X and Y , and write $Z = X \cdot Y$ (or $Z = X \times Y$), see Fig. 2.2.

Braids were explicitly introduced by E. Artin [16] in 1925, although they had already been used implicitly in 1891. Since then, they have proven to be an important concept, in fact, there is a whole branch of mathematics, called *braid theory*, which has applications in many areas of pure mathematics, and in applied areas such as fluid mechanics. It is therefore important to find a convenient representation of braids, which is well suited for calculations.

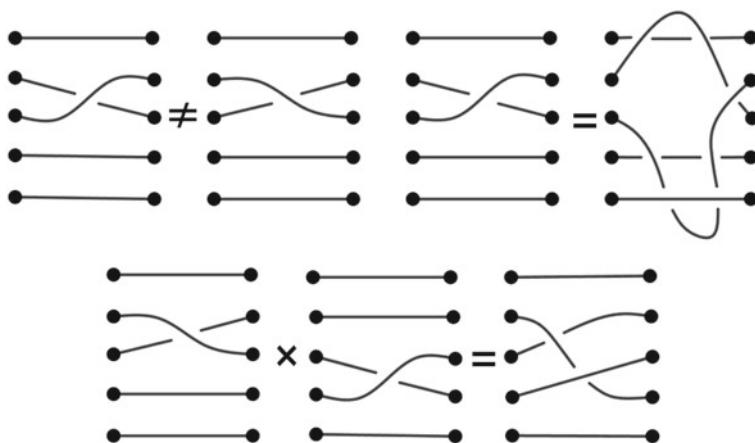


Fig. 2.2 Braids and their products

Plane Transformations and 2×2 Matrices

To see what we mean by a “representation suitable for calculations”, let us consider an example from geometry, namely, plane rotations. Any such rotation fixes some point O in the plane, selects some angle α , and direction (clockwise or counter-clockwise), and then “rotates” the plane around O by the angle α in a given direction. To represent such a rotation algebraically, we assume that our plane is the coordinate plane with centre O . Then the image of a point with coordinates (x, y) after counter-clockwise rotation by the angle α is a point with coordinates $(x \cos \alpha - y \sin \alpha, x \sin \alpha + y \cos \alpha)$.

In general, any transformation of the plane sending the point (x, y) to the point $(ax + by, cx + dy)$ is called a *linear* transformation. Any linear transformation is fully determined by the coefficients a, b, c, d , and it is convenient to write these coefficients in the form of a 2×2 table $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, which is called a 2×2 *matrix*.

If we rotate the plane counter-clockwise first by the angle α and then by the angle β , the “total” rotation is by angle $\alpha + \beta$, hence the point (x, y) is sent to $(x \cos(\alpha + \beta) - y \sin(\alpha + \beta), x \sin(\alpha + \beta) + y \cos(\alpha + \beta))$. More generally, if we first apply the linear transformation $(x, y) \rightarrow (a_1x + b_1y, c_1x + d_1y)$, and then the transformation $(x, y) \rightarrow (a_2x + b_2y, c_2x + d_2y)$, then the image of the point (x, y) has coordinates

$$a_2(a_1x + b_1y) + b_2(c_1x + d_1y) = (a_2a_1 + b_2c_1)x + (a_2b_1 + b_2d_1)y$$

and

$$c_2(a_1x + b_1y) + d_2(c_1x + d_1y) = (c_2a_1 + d_2c_1)x + (c_2b_1 + d_2d_1)y.$$

The corresponding matrix $\begin{bmatrix} a_2a_1 + b_2c_1 & a_2b_1 + b_2d_1 \\ c_2a_1 + d_2c_1 & c_2b_1 + d_2d_1 \end{bmatrix}$ is called the *product* of the matrices $\begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix}$ and $\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix}$. For example,

$$\begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 3 \cdot 0 + 0 \cdot 1 & 3 \cdot 1 + 0 \cdot 0 \\ 0 \cdot 0 + 2 \cdot 1 & 0 \cdot 1 + 2 \cdot 0 \end{bmatrix} = \begin{bmatrix} 0 & 3 \\ 2 & 0 \end{bmatrix}.$$

$n \times n$ Matrices and Their Products

Similarly, any rotation in a 3-dimensional space sends the point (x, y, z) to the point $(ax + by + cz, dx + ey + fz, gx + hy + iz)$, where the coefficients $a, b, c, d, e, f, g, h, i$ depend on the angles of rotation. It is convenient to write these coefficients in a 3×3 table. In general, a square matrix is an $n \times n$ array of elements of the form

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}.$$

The product of $n \times n$ matrices A and B is an $n \times n$ matrix D with elements $d_{ij} = \sum_{k=1}^n a_{ik}b_{kj}, \forall i, j$. This definition of product generalizes the product of 2×2 matrices defined above, and is designed in such a way that the composition of rotations (or any other linear transformations of the space) corresponds to the product of the corresponding matrices.

Representation of rotations/transformations in terms of matrices allows us to forget about geometry, and we can answer essentially all interesting questions about rotations/transformations by working with matrices only. In the case of relatively simple objects like two-dimensional rotations this may be not very beneficial, but being able to represent complicated objects like braids in terms of matrices would make the computations relatively effortless, because matrices are much easier to analyse.

Encoding Braids as Matrices

Elements of matrices can be rational numbers, real numbers, and, more generally, they may belong to any set F for which the basic operations $(+, -, \cdot, /)$ are well-defined, that is, to any *field* F , see Sect. 1.7 for the exact definition and more examples of fields. Matrices whose elements belong to a field F are called matrices *over* the field F .

One way to encode braids as matrices over some field had already been suggested by Burau in the 1930s, and it works well for braids with $m \leq 3$ strands. However, it was later shown [54] that for braids with $m \geq 5$ strands some different braids correspond to the same matrix. Finally, Bigelow [55] and (independently and using different methods) Krammer [232] developed a way to represent braids as matrices without this disadvantage.

Theorem 2.2 *For every m , there exist a natural number n , a field K , and a map ρ assigning to any braid X with m strands an $n \times n$ matrix $\rho(X)$ over the field K , such that*

- (a) $\rho(X) \neq \rho(Y)$ whenever $X \neq Y$ (different matrices correspond to different braids), and
- (b) $\rho(X \cdot Y) = \rho(X) \cdot \rho(Y)$ for any two braids X, Y .

In the language of group theory, Theorem 2.2 states that “braid groups are linear”. Essentially, it reduces the study of braids (and their products) to the study of matrices (and their products). This is very useful, because a lot of tools developed to study matrices can now be applied to braids.

Reference

D. Krammer, Braid groups are linear, *Annals of Mathematics* **155**-1, (2002), 131–156.

2.3 Explicit Expander Constructions Using the Zig-Zag Product

Which Computer Networks are Reliably Connected?

Imagine a large number of computers which need to be connected into a single network. We could just connect any two by a direct cable, but this is too expensive. Another option is to enumerate all computers from 1 to n , and connect computers i and $i + 1$, $i = 1, 2, \dots, n - 1$. This is much cheaper, but if just one cable is broken, some subset S of the computers will be disconnected from the others. If S consists of a small number of computers, this is a minor issue, but, in the worst case, if n is even and the cable between $n/2$ and $n/2 + 1$ is broken, the set $S = \{1, 2, \dots, n/2\}$ consisting of $|S| = n/2$ computers will be disconnected.

Intuitively, the quality of any network G is high if a small number of broken cables cannot disconnect a large set of computers from the others. For any subset S of computers, denote by $|\partial S|$ the number of cables directly connecting computers from S to computers from outside S . If there is an adversary who wants to break the network, he can break $|\partial S|$ cables to disconnect $|S|$ computers. If he wants to break a few cables and disconnect a lot of computers, he will select a set S with break-to-disconnect ratio $|\partial S|/|S|$ as low as possible. Hence, the quality of G can be measured by the parameter

$$h(G) = \min_{1 \leq |S| \leq n/2} \frac{|\partial S|}{|S|}.$$

For example, in a network with quality $h(G) = 0.01$ it may be possible to break k cables and disconnect $100k$ computers, while in any network with $h(G) \geq 0.5$ any k broken cables will disconnect at most $2k$ computers, which is much less critical.

For the linear connection described above, $h = 2/n$, so the quality goes to 0 as $n \rightarrow \infty$. In contrast, if G_n is a network with n computers in which every two are directly connected by a separate cable (such a network is called *complete*), then $|\partial S| = |S|(n - |S|)$, and $h(G_n) = \min_{1 \leq |S| \leq n/2} \frac{|S|(n - |S|)}{|S|} = \min_{1 \leq |S| \leq n/2} (n - |S|) = n/2$, which is excellent, but, as we have mentioned, too expensive. The question is, can we build networks with reasonable quality, but with as few cables as possible?

Graphs, Regular Graphs, and Families of Expanders

The standard mathematical model for a computer network is a *graph*. A graph G of size n is a set of n vertices, some pairs of which are connected by edges. In our example, vertices are computers, and edges are cables, but graphs have many other applications. For example, vertices may be cities and edges high-quality highways to be built between them. In this case, a low parameter $h(G)$ implies that there is a large group S of cities connected to others by just a few highways, which may be problematic in terms of traffic on these highways. So, once again, it would be nice to build highways in such a way that $h(G)$ is high, but the total number of highways is reasonable.

A graph G with high $h(G)$ is called a good *expander*. However, constructing *one* good expander of size n would not help to build a network with $k > n$ computers. An infinite family of graphs $G_1, G_2, \dots, G_i, \dots$ is called a *family of expanders* if there is an $\varepsilon > 0$ such that $h(G_i) \geq \varepsilon$ for all i . A family of expanders can help to build networks with quality at least ε , no matter how many computers we have. To bound the number of cables/edges, a standard approach is to require that every vertex is connected to exactly d others, where d is a fixed number such as $d = 4$. Such graphs are called *d -regular*.

Random and Non-random Methods to Produce Expanders

Do there exist d -regular families of expanders, which would help to build relatively cheap but high quality networks? For $d \geq 3$, the answer is yes, and, moreover, almost every network would work! That is, if we connect any computer to d other computers chosen at random, then, with high probability, the resulting network will have a high quality. This method, however, is not practical, because building large networks is too expensive to rely on chance. What if we are unlucky and produce a low-quality network? We therefore need a *deterministic*, that is, non-random method to produce d -regular families of expanders.

Several such methods have been suggested starting from 1973, but the proofs are complicated, and it is hard to intuitively understand why the constructed graphs are expanders. In contrast, Reingold, Vadhan and Wigderson [319] constructed a much simpler family of expander graphs, $G_1, G_2, \dots, G_i, \dots$, based on the notion of *zig-zag product* they introduced. The zig-zag product is a smart way to take a d_1 -regular graph G of size n and a d_2 -regular graph H of size d_1 , and produce a new d_2^2 -regular graph R of larger size $n \cdot d_1$, in such a way that if both G and H are good expanders, then so is R . They denote the graph R by $G \cdot_z H$ and call it the *zig-zag product* of G and H . With the zig-zag product, one can start with some good expanders G_0 and H , and then construct $G_1 = G_0 \cdot_z H, G_2 = G_1 \cdot_z H$, and so on, building larger and larger good expanders in a deterministic way.

The Zig-Zag Product

To explain the definition of the zig-zag product, we assume that edges of the graphs G and H are coloured by d_1 and d_2 colours, respectively, such that all edges sharing the same vertex have a different colour. We also colour the vertices of the graph H using the same d_1 colours as the edges of the graph G . For example, let G be a 4-regular graph with 6 vertices, such that vertex 1 is connected with, say, vertices 2, 3, 5, and 6, using edges coloured in pink (p), orange (o), green (g), and yellow (y), respectively. Let H be a cycle of length 4, that is, a graph with vertices p, o, g , and y (note that the vertices of H correspond to the edge colours of G), and edges $p - o$, $o - g$, $g - y$, and $y - p$. Then we colour the edges of H with two new colours, say, red and blue, so that, say, edges $p - o$ and $g - y$ are red (r), while edges $o - g$ and $y - p$ are blue (b).

Now, the graph $G \cdot_Z H$ has $n \cdot d_1$ vertices, which can be visualized as points in an $n \times d_1$ grid. The rows of the grid corresponds to vertices of H (or, equivalently, to colours of edges in G), while the columns in the grid corresponds to vertices of G . In our example, $G \cdot_Z H$ has 24 vertices, which forms a grid of size 6×4 . The vertex in row x and column i will be denoted x_i , see Fig. 2.3.

Next, take any vertex v of $G \cdot_Z H$, say, $v = g_1$. It corresponds to vertex 1 in G and vertex g in H . Also, select any two colours used for the edges of the graph H (order matters, repetition is allowed, so we can choose red-red, red-blue, blue-red, or blue-blue, let us choose red-blue). We describe an edge going from $v = g_1$ and coloured “red-blue” using the following 3-step procedure.

- Step “Zig”: Go from vertex g in H along the edge coloured red, and arrive at vertex y – yellow.
- Step “Entropy wave” (all step names are suggested by the authors): Go from vertex 1 in G along the yellow edge and arrive at vertex 6, see Fig. 2.3.
- Step “Zag”: Go from vertex y in H along the edge coloured blue, and arrive at vertex p – pink.

Hence, the vertex g_1 in $G \cdot_Z H$ should be connected by a red-blue edge to the vertex p_6 . In a similar way, the combinations red-red, blue-red, and blue-blue lead to three more edges going from g_1 , and, similarly, from any other vertex, resulting in a 4-regular graph with 24 vertices.

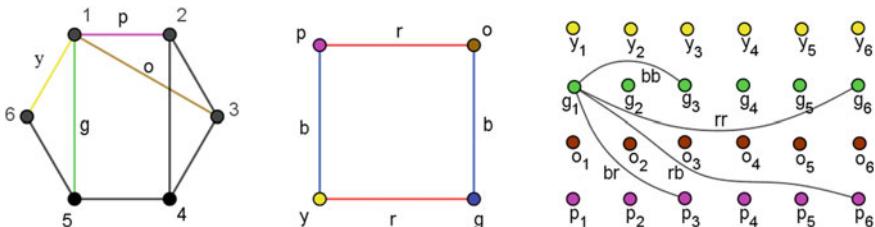


Fig. 2.3 An example of a zig-zag product

An Explicit Family of Expanders

There is one more graph operation which preserves (and even improves) expansion – taking a “square”. For any graph G , a 2-edge path is a set of vertices i, k, j , connected by edges $i - k$ and $k - j$. Now, the graph G^2 is the graph with the same vertices as G such that for every 2-edge path i, k, j in G , we put an edge between i and j in G^2 . If G is d -regular, then there are d^2 2-edge paths starting from any vertex i in G , therefore G^2 is d^2 -regular.

With these operations in hand, Reingold, Vadhan and Wigderson start with any *one* d -regular graph H of size d^4 and sufficiently high $h(H)$, and then construct a family of expanders iteratively, using the formulas $G_1 = H^2$, $G_{i+1} = G_i^2 \cdot_Z H$, $i \geq 1$. For every i , G_i is a d^2 -regular graph of size d^{4i} .

Theorem 2.3 $G_1, G_2, \dots, G_i, \dots$ is a d^2 -regular family of expanders.

The main part in proving Theorem 2.3 in [319] is to show that if G and H are good expanders, then so is $G \cdot_Z H$, and the rest follows easily. However, the main contribution of the authors is not the proof of Theorem 2.3, but the construction of the zig-zag product itself.

An Improved Zig-Zag Product

The authors also improved the construction in different ways.

First, the time to actually build any graph G_i cannot be less than its number of vertices, d^{4i} . However, as i grows, d^{4i} quickly becomes huge, so the whole graph would be difficult to describe even with the fastest computers we have nowadays. The authors modified the construction so that the resulting expanders can be constructed locally. That is, you do not need to compute the whole huge graph, you only choose a single vertex $v \in G_i$ and there is an algorithm which tells you all the edges going from this vertex. This is actually all you need in many applications, and it can be done in time $P(\ln(d^{4i})) = P(4i \ln d)$, where P is some polynomial. For reasonably large i , this is *much* faster than d^{4i} .

Second, if the initial graph H is, say, 30-regular, then the resulting family of expanders is 900-regular. For many applications, this is too much. The authors describe a method which uses Theorem 2.3 to build a d -regular family of expanders with $d = 4$, and even with $d = 3$.

The zig-zag product provides an explicit and fast method to build stable networks of computers of any size, to plan highways between cities avoiding significant traffic problems, etc.

Application: Finding Your Hotel in a Town

Expander graphs play a major role in many areas of computer science and mathematics. In many applications, it is crucial to have an efficient and deterministic method for constructing such graphs. For these reasons, it is not surprising that many new

interesting and important theorems have been and will be proved using the concept of zig-zag product and Theorem 2.3. Here, we briefly mention just one application.

Imagine that you are travelling, spend a night in a hotel, but then go for a walk and get lost in a town. The bad news is that you have no map, do not know the address of the hotel, and people in the town do not speak your language. The good news is that you remember what your hotel looks like, you remember that it is located near a crossroads, and there are not so many (around a hundred) crossroads in the town. It takes you about 2–3 minutes to walk from one crossroad to an adjacent one, so you can just check them all in about 3–5 hours.

However, how exactly should you organize your search? If you just walk at random, there is a non-zero chance that you may be unlucky and, while visiting other crossroads multiple times, never visit the crossroads with your hotel. In order to not rely on chance, you may want to develop a systematic and deterministic strategy for your search. There are many easy algorithms for doing this (the most well-known of them are called “breadth-first search” and “depth-first search” algorithms), but they require you to *memorize* all the crossroads you have visited before to avoid visiting them again and again. As you can imagine, such memorising can be very challenging. Is there a method to find your hotel which does not require you to memorize too much information on the way?

Formally, we can interpret crossroads as vertices of a graph, and streets as edges, and then the question is: given a graph G formed of n vertices, and two vertices s and t in G , decide whether s and t are connected by a path in G , and if so, find such a path. Using the concept of zig-zag product, Reingold [318] developed a deterministic algorithm which solves this problem using at most $C \ln n$ memory, where C is some constant. For large n , $C \ln n$ is much-much lower than n , so this algorithm may be helpful for finding a hotel. It also has numerous other applications, and is of fundamental theoretical importance.

Reference

O. Reingold, S. Vadhan and A. Wigderson, Entropy waves, the zig-zag graph product, and new constant-degree expanders, *Annals of Mathematics* **155**-1, (2002), 157–187.

2.4 Elliptic Curves Over Function Fields Can Have Arbitrarily Large Rank

Integer and Rational Solutions to Linear and Quadratic Equations

One of the oldest topics in mathematics is to find integer solutions to equations. The simplest equations are linear, in the form $ax + by = c$ for some integers a, b, c , and they are very easy to solve. Consider, for example, the equation $2x - 3y = 7$. We can find x as $x = \frac{7+3y}{2}$. If y is even, then x is not an integer. Hence, y is odd and can be written as $y = 2k + 1$ for integer k . Then $x = \frac{7+3(2k+1)}{2} = 3k + 5$. The general case can be solved similarly.

A bit more interesting are quadratic equations. For example, the question “What are the right triangles with integer sides” reduces to finding integer solutions to the equation $a^2 + b^2 = c^2$. Dividing it by c^2 , we get $(a/c)^2 + (b/c)^2 = 1$, and the question reduces to finding *rational* solutions to the equation $x^2 + y^2 = 1$. Geometrically, this equation describes a circle in the coordinate plane, and we are interested in finding points with rational coefficients (which are called just “rational points”) on this circle. Let us take any one such point, say, $X_0 = (-1, 0)$, and draw any line through it with rational slope. The general equation of a line is $y = kx + b$. Because X_0 is on the line, we have $-1 = k \cdot 0 + b$, hence $b = -1$, and the equation of the line is $y = kx - 1$. Let us find another point where this line intersects the circle $x^2 + y^2 = 1$. The x -coordinate of this point satisfies the equation $x^2 + (kx - 1)^2 = 1$, or $x^2 + k^2x^2 - 2kx = 0$. The roots are $x = 0$, and $x = \frac{2k}{1+k^2}$. Because k is rational, we can write $k = \frac{n}{m}$ for integers n, m , then $x = \frac{2k}{1+k^2} = \frac{2mn}{n^2+m^2}$, and $y = kx - 1 = \frac{n}{m} \frac{2mn}{n^2+m^2} - 1 = \frac{n^2-m^2}{n^2+m^2}$.

Hence, integer solutions to the original equation $a^2 + b^2 = c^2$ are values a, b, c such that $\frac{a}{c} = x = \frac{2mn}{n^2+m^2}$ and $\frac{b}{c} = y = \frac{n^2-m^2}{n^2+m^2}$, which implies that $a = 2mn$, $b = n^2 - m^2$, $c = n^2 + m^2$, where m and n are arbitrary integers. For example, $n = 2$, $m = 1$, gives $a = 4$, $b = 3$, $c = 5$, and one can check that indeed $3^2 + 4^2 = 5^2$. Again, the rational solutions of a general quadratic equation of the form $ax^2 + bxy + cy^2 + dx + ey + f = 0$ can be found similarly: take any one solution, draw all possible lines with rational slopes thought it, and all the solutions are intersections of the lines with the corresponding quadratic curves.

Rational Solutions to Cubic Equations

The next case is the cubic equation of the form

$$Ax^3 + Bx^2y + Cxy^2 + Dy^3 + Ex^2 + Fxy + Gy^2 + Hx + Iy + J = 0,$$

with rational coefficients. With an appropriate change of variables, it can be reduced to the much simpler form

$$y^2 = x^3 + ax + b. \quad (2.5)$$

The set of all real solutions to (2.5) is called an *elliptic curve*, and understanding the structure of rational points on an elliptic curve is one of the central questions in modern mathematics.

The simplest observation is that if (x_0, y_0) is a rational point, then so is its reflection $(x_0, -y_0)$. Indeed, if (x_0, y_0) is a rational point on the elliptic curve (2.5), then $(y_0)^2 = (x_0)^3 + a(x_0) + b$. But then $(-y_0)^2 = (x_0)^3 + a(x_0) + b$, hence $(x_0, -y_0)$ is a rational point on the curve as well.

More interestingly, if (x_1, y_1) and (x_2, y_2) are two rational points with $x_1 \neq x_2$, then, if we draw a line through them, it intersects the curve (2.5) in a third rational point. Indeed, if the equation of the line is $y = kx + c$ with rational k, c , then (2.5)

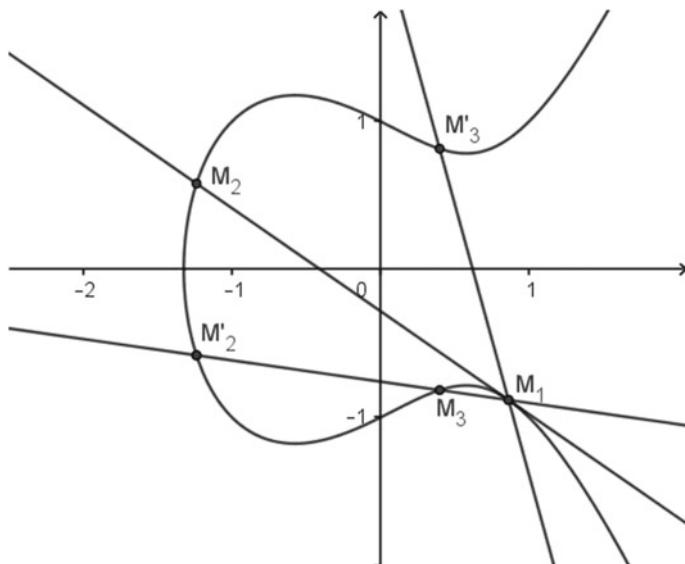


Fig. 2.4 A sequence of rational points on an elliptic curve

implies $x^3 + ax + b - (kx + c)^2 = 0$. It is known that sum of all three solutions of cubic equation with leading coefficient 1 is equal to the coefficient of x^2 with opposite sign, in our case k^2 , hence the third solution x_3 is equal to $k^2 - x_1 - x_2$. Because k, x_1, x_2 are rational numbers, and sums and products of rational numbers are rational numbers, $x_3 = k^2 - x_1 - x_2$ is a rational number, hence so is $y_3 = kx_3 + c$.

If x_2 is moving on the elliptic curve toward x_1 , the line through (x_1, y_1) and (x_2, y_2) becomes a tangent line through (x_1, y_1) , and the same argument as above implies that, if this line is not vertical, it intersects the curve in another rational point. Hence, we can start with any *one* rational point $M_1 = (x_1, y_1)$, find another one $M_2 = (x_2, y_2)$ using the tangent line, then another one $M'_2 = (x_2, -y_2)$ using reflection. Then the line passing through M_1 and M'_2 intersects the elliptic curve at another rational point $M_3 = (x_3, y_3)$, and one more rational point is its reflection, $M'_3 = (x_3, -y_3)$, see Fig. 2.4. If we continue this process, we can either return back to the initial point $M_1 = (x_1, y_1)$ after a finite number of steps, or generate infinitely many rational points. In the latter case, the initial point $M_1 = (x_1, y_1)$ is said to be of *infinite order*.

The Rank of an Elliptic Curve

It might be that not all rational points of (2.5) are generated using just one initial point, but there is a fundamental theorem stating that there always exists a *finite set* S of rational points which generate all others in the way described above. Obviously, we can assume that no points in S can be generated by others, otherwise such points could be excluded. Such a set S is called a generating set of rational points of the elliptic

curve. The number of points of infinite order in S is called the *rank* of the elliptic curve (2.5). In fact, the same elliptic curve may have many different generating sets. However, one can prove that any two such sets always contain the same number of points of infinite order, so that the rank of an elliptic curve depends only on the curve, but not on the particular choice of S .

If we could explicitly write down any “generating set” S of (2.5), we would describe *all* rational solutions to (2.5). However, this turns out to be very difficult. Even understanding how many points we need, that is, what the rank of the elliptic curve is, is already a very deep problem. There is a convenient formula for calculating the rank of any elliptic curve, but nobody can prove that it is correct. The correctness of that formula is known as the “Birch and Swinnerton-Dyer conjecture”. At the beginning of the 21st century, the Clay Mathematics Institute identified [85] seven problems, in their opinion the most important and challenging, and for each problem offered a million-dollar prize for its solution. The Birch and Swinnerton-Dyer conjecture is one of the problems in this list. In fact, it is hard even to understand how large the rank of an elliptic curve can be. By 2002, it was known [262] that there is a curve with rank at least 24, but can it be 30, 100, 1000? Nobody knows.

Elliptic Curves Over Arbitrary Fields

Equation (2.5) can be studied not only for rational numbers, but for real numbers, complex numbers, and, more generally, for any abstract sets, for which the usual arithmetic operations (addition, subtraction, multiplication, and division) are well defined and satisfy the usual properties, such as $a(b + c) = ab + ac$, etc. Such sets are called *fields*. For example, for any prime p , we can build the field \mathbb{F}_p consisting of elements $0, 1, 2, \dots, p - 1$ with operations defined “modulo p ”, so that $(p - 1) + 1 = p = 0$, $(p - 1) + 2 = p + 1 = 1, 2(p - 1) = 2p - 2 = p - 2$, hence $(p - 2)/(p - 1) = 2$, and so on, see Sect. 1.7 for more details. We can also define the field $\mathbb{F}_p(t)$ of rational functions of the form

$$f(t) = \frac{a_m t^m + a_{m-1} t^{m-1} + \dots + a_1 t + a_0}{b_n t^n + b_{n-1} t^{n-1} + \dots + b_1 t + b_0},$$

where $a_m, \dots, a_1, a_0, b_n, \dots, b_1, b_0 \in \mathbb{F}_p$. Then Eq. (2.5) becomes

$$y(t)^2 = x(t)^3 + a(t)x(t) + b(t), \quad (2.6)$$

where $a(t), b(t) \in \mathbb{F}_p(t)$ are given functions, and $x(t), y(t) \in \mathbb{F}_p(t)$ are unknown functions. For example, the equation $y(t)^2 = x(t)^3 + x(t)/t + 4t$ in $\mathbb{F}_5(t)$ has a solution $y(t) = t^3$, $x(t) = t^2$, because $(t^2)^3 + t^2/t + 4t = t^6 + 5t = t^6 + 0t = (t^3)^2$. The rank of (2.6) can be defined in a similar way, and, surprisingly, it is somewhat easier to understand than for the case of rational solutions. In particular, Douglas Ulmer [382] proved the following theorem.

Theorem 2.4 *Let E be the elliptic curve defined over the field $\mathbb{F}_p(t)$ by the equation*

$$y^2 + xy = x^3 - t^d,$$

where $d = p^n + 1$ for some positive integer n . Then the rank of the curve E is at least $(p^n - 1)/2n$.

Theorem 2.4 implies that the rank of an elliptic curve over the field $\mathbb{F}_p(t)$ can be as large as we want. This result is not known to hold for the “usual” elliptic curves (2.5). In fact, many experts today believe that the rank of the “usual” elliptic curves with rational coefficients is in fact bounded above by an absolute constant. Hence, Theorem 2.4 provides evidence for the fact that elliptic curves over function fields may behave in special ways which have no counterpart in the most important case, namely, the field of rational numbers.

Reference

D. Ulmer, Elliptic curves with large rank over function fields, *Annals of Mathematics* **155**-1, (2002), 295–315.

2.5 The Optimality of the Standard Double Bubble

How to Build a City of a Given Area and Minimal Perimeter

When ancient people came to new land to build their cities, they needed to surround them by a wall to defend from adversaries. Assuming that a city should have area at least one square kilometer to accommodate all the people, what is the minimum length of wall they need to build? Obviously, they could build a city in the form of a square with side length 1 km, and then the length of the wall is 4 km in total, but could they do better? Forming a rectangle with sides x and $1/x$ does not help. Indeed, the perimeter of such a rectangle is $P(x) = 2(x + 1/x)$, and we need to choose x so that $P(x)$ is minimal. If you know the concept of the *derivative*, you can check that $P'(x) = 2 - 2/x^2$, hence $P'(x) = 0$ if $x = 1$. But the case $x = 1$ corresponds to the square, hence no rectangle with area 1 has smaller perimeter than the unit square. If you are not familiar with differentiation, you can deduce the inequality $2(x + 1/x) \geq 4$ from the fact that $(\sqrt{x} - 1/\sqrt{x})^2 \geq 0$.

Another option is to form a circle of radius r with area $\pi r^2 = 1$, hence $r = 1/\sqrt{\pi}$, and the wall length is $2\pi r = 2\sqrt{\pi} \approx 3.54 < 4$. It turns out than this is optimal, and this fact was already known in Ancient Greece. Investigations show that many ancient cities were indeed built in the form of circles.

How to Build Two Cities of Equal Areas and Minimal Perimeter

Now, what if two different groups of people decided to build their cities of area 1 km^2 each, and agree to share a common wall to minimize effort? For each city, the optimal form is a circle, but two circles cannot benefit from sharing a wall, so people would need to build $2(2\sqrt{\pi}) = 4\sqrt{\pi} \approx 7.09$ km of wall in total. In contrast,

two squares of side length 1 can share one side, hence only 7 sides need to be built, and $7 < 4\sqrt{\pi}$. Moreover, in this case squares can be beaten by rectangles with sides x and $1/x$, sharing the side of length x . Indeed, the total wall length is then $f(x) = 3x + 4(1/x)$. To find x such that $f(x)$ is minimal, we find $f'(x) = 3 - 4/x^2$, and solve the equation $f'(x) = 0$, resulting in $x = \sqrt{4/3} \approx 1.15$, and the total wall length is $f(\sqrt{4/3}) = 4\sqrt{3} \approx 6.93 < 7$. Without differentiation, you can deduce the inequality $3x + 4(1/x) \geq 4\sqrt{3}$ directly from the fact that $(\sqrt{3x} - \sqrt{4/x})^2 \geq 0$. Intuitively, rectangles beat squares because in this case cities can share a longer wall.

How can we do better? Maybe we could first build a large city with area 2 km^2 , and then divide it by half? The large city will have a minimal external wall if it is a circle of radius r such that $\pi r^2 = 2$, hence $r = \sqrt{2/\pi}$. After this, we need to separate the cities by an internal wall of length $2r$, therefore the total length of the wall is $2\pi r + 2r \approx 6.61 < 6.93$.

Is this optimal or can we do even better? In this case, the optimal configuration was not known to ancient people, and remained an open question until the problem was solved [153] by a team of undergraduate students in 1993. The optimal construction turns out to be cities in the form of two circles, intersecting at points A and B , and separated by a straight common wall AB , such that at each of points A and B three meeting walls divide the full angle of 360° into three equal angles of 120° each, see Fig. 2.5. To make the areas of such cities equal to 1, the radii of the circles should be about $r \approx 0.629$. The wall consists of two circular parts of length $2/3(2\pi r)$ each, plus a straight segment AB of length $\sqrt{3}r$, totalling $(8/3)\pi r + \sqrt{3}r \approx 6.36 < 6.61$. The team of students proved that this is optimal.

How to Build Two Cities of Given Areas and Minimal Perimeter

Obviously, there is no reason why the cities should have equal areas. One group of people can be larger and would like to live in a larger city. It is therefore natural to ask how we would build a wall of minimal length surrounding two cities with areas S_1 and S_2 . The undergraduates solved this case as well, and the optimal construction

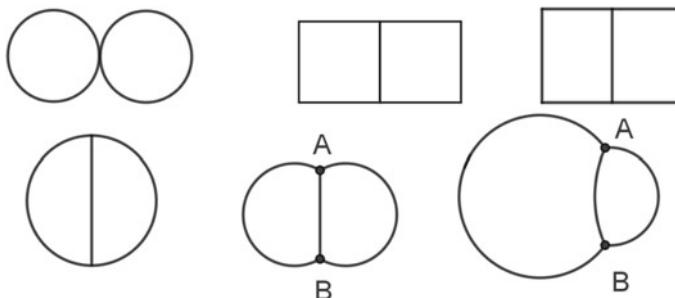


Fig. 2.5 Some ways to build two cities with a common wall

is similar. This time the city walls are circles of *different* radii r_1 and r_2 , intersecting at points A and B , but the internal wall AB is now in the form of another arc, such that at each of the points A and B the three meeting arcs divide the full angle of 360° into three equal angles of 120° each, as previously, see Fig. 2.5. Now, we can express the areas of the two resulting cities as some functions $f(r_1, r_2)$ and $g(r_1, r_2)$, respectively, and then find r_1 and r_2 from the system of equations $f(r_1, r_2) = S_1$ and $g(r_1, r_2) = S_2$. This way we get cities of the required areas, and the total length of the wall turned out to be minimal. There are, however, no known pairs of ancient cities connected in this form, because people from that time did not know the solution.

Why a Soap Bubble Has the Form of a Sphere

A similar problem arises in 3-dimensional space even more naturally. Usually soap bubbles have a fixed amount of air inside them, and therefore have a constant volume. Also, physical laws imply that a bubble will try to make its surface area as small as possible. So, a bubble is trying to “solve” a 3-dimensional version of our wall problem above, namely, to find a shape with a given volume (say, 1) and minimal surface area. For example, the cube with edge length 1 has surface area 6, but we never see cubical bubbles, they are usually in the form of a sphere. The volume of a sphere of radius r is $V = \frac{4}{3}\pi r^3$, and is equal to 1 if $r = \sqrt[3]{3/4\pi}$. The surface area is given by the formula $S = 4\pi r^2$, and is equal to $4\pi(\sqrt[3]{3/4\pi})^2 \approx 4.84 < 6$. The fact that this is optimal was conjectured by Archimedes, but rigorously proved only by Schwarz [341] in 1884.

How Two Equal Soap Bubbles Connect Together

In practice, two bubbles often collide, and form a “double bubble”, that is, two volumes V_1 and V_2 connected together. What is the minimal surface area needed to enclose and separate these volumes? Again, let us start with some experiments in the case $V_1 = V_2 = 1$. Two spheres can touch at most in one point, hence their surface area is about $2 \cdot 4.84 = 9.68$. Two unit cubes can share a face, so 11 faces are needed, with total area $11 > 9.68$, hence, while in the 2-dimensional case squares beat circles, in three dimensions cubes are even worse than spheres. Two semi-spheres form a sphere of volume 2 if their radii are such that $\frac{4}{3}\pi r^3 = 2$, $r = \sqrt[3]{3/2\pi}$, the surface area of a sphere is $4\pi r^2$, and the separating surface is a circle with area πr^2 , resulting in a total area of $5\pi(\sqrt[3]{3/2\pi})^2 \approx 9.60 < 9.68$. The team of undergraduates who solved the 2-dimensional case conjectured that the optimal construction in three dimensions is similar: two spheres, intersecting at a circle C , with the separating surface being the flat disk with boundary C . In 1995, a team of researchers proved [190] that this is optimal by reducing the problem to calculating 200260 integrals, and then using a computer to perform the calculation. However, this method worked only if $V_1 = V_2$.

How Two Non-equal Soap Bubbles Connect Together

For $V_1 \neq V_2$ the conjectural construction was also similar to the 2-dimensional case: two spheres with radii r_1 and r_2 , respectively, intersecting at a circle C , and a separating surface in the form of another sphere going through C , such that the three spheres, while intersecting, form three equal angles of 120° each. Finally, r_1 and r_2 are chosen so that the volumes of the resulting parts are V_1 and V_2 . This construction is known as the *standard double bubble*.

The following theorem of Michael Hutchings, Frank Morgan, Manuel Ritoré and Antonio Ros [207] shows that the standard double bubble is indeed optimal.

Theorem 2.5 *In \mathbb{R}^3 , the unique perimeter-minimizing double bubble enclosing and separating regions R_1 and R_2 of prescribed volumes V_1 and V_2 is a standard double bubble, consisting of three spherical caps meeting along a common circle at 120-degree angles. (For equal volumes, the middle cap is a flat disc.)*

In contrast to the earlier proof of the $V_1 = V_2$ case, the proof of Theorem 2.5 is computer-free.

Bubbles cannot prove theorems, but they “knew” the result well before it was proved, and always form the standard double bubble after collisions. Moreover, they form a (conjecturally) optimal shape even if three or more bubbles collide at once, while the mathematical theorem confirming that this shape is optimal has not yet been proved.

Reference

M. Hutchings, F. Morgan, M. Ritoré and A. Ros, Proof of the double bubble conjecture, *Annals of Mathematics* **155**-2, (2002), 459–489.

2.6 Counting Integer Solutions of Equations in Three Variables

Integer Solutions of Homogeneous Polynomial Equations

One of the oldest problems in mathematics is finding integer solutions of polynomial equations, or at least estimating how many solutions an equation can have. A *monomial* in n variables x_1, \dots, x_n of degree d is a function of the form $f(x_1, \dots, x_n) = ax_1^{k_1}x_2^{k_2}\dots x_n^{k_n}$, where $a \in \mathbb{Z}$ (\mathbb{Z} denotes the set of integers), and k_1, \dots, k_n are non-negative integers with $\sum_{i=1}^n k_i = d$. A homogeneous polynomial F of degree d , or *form* of degree d , is a sum of any number of such monomials. For example, $2x^3yz$ and $-3xy^2z^2$ are monomials in $n = 3$ variables of degree $d = 5$, and $2x^3yz - 3xy^2z^2$ is an example of a form of degree 5. On the other hand, $2x^3yz + xyz$ is not a form, because the degrees of the monomials $2x^3yz$ and xyz are not the same.

Given any form F in n variables, how many integer solutions can the equation

$$F(x_1, x_2, \dots, x_n) = 0 \quad (2.7)$$

have?

The Cases of $n = 1$ and $n = 2$ Variables

If $n = 1$, the only form of degree d is $F(x) = ax^d$, $a \neq 0$, and the equation $ax^d = 0$ has solution $x = 0$.

Let $n = 2$, and let us start with the easy equation $x^2 - 3xy + 2y^2 = 0$. If $y \neq 0$, we can divide both sides of the equation by y^2 , to get $(x/y)^2 - 3(x/y) + 2 = 0$, or $z^2 - 3z + 2 = 0$, where $z = x/y$ is a new variable. This equation has roots $z_1 = 1$ and $z_2 = 2$, hence either $x/y = 1$ or $x/y = 2$. This results in two infinite families of integer solutions, the first one is $(x, y) = (k, k)$, $k \in \mathbb{Z}$, and the second one is $(x, y) = (2k, k)$, $k \in \mathbb{Z}$.

In general, the solution method is similar. We have an equation of the form

$$a_0x^d + a_1x^{d-1}y + \dots + a_{d-1}xy^{d-1} + a_dy^d = 0, \quad (2.8)$$

where at least one coefficient is non-zero. If $y = 0$, then (i) if $a_0 = 0$, then x can be arbitrary, (ii) if $a_0 \neq 0$, then $x = 0$. If $y \neq 0$, let us divide both sides of the equation by y^d , and introduce a new variable $z = x/y$ to get the equation

$$a_0z^d + a_1z^{d-1} + \dots + a_{d-1}z + a_d = 0. \quad (2.9)$$

Equation (2.9) has at most d solutions, let m of them be rational, and let $z_1 = \frac{a_1}{b_1}$, $z_2 = \frac{a_2}{b_2}, \dots, z_m = \frac{a_m}{b_m}$ be their representation in the form of irreducible fractions such that $a_i \geq 0$, $i = 1, \dots, m$. Then x and y satisfy one of the equations $\frac{x}{y} = \frac{a_i}{b_i}$, $1 \leq i \leq m$, each producing an infinite family solutions of the form $x = a_i k$, $y = b_i k$, $k \in \mathbb{Z}$.

Counting Simple Solutions

Let us call a solution (x, y) to (2.8) *simple* if x and y have no common factor, and either $x > 0$, or $x = 0$ and $y > 0$. If we can find all simple solutions of some equation, all the other solutions can be produced from simple ones by scaling. For example, the equation $x^2 - 3xy + 2y^2 = 0$ has two simple solutions: $x = y = 1$ and $x = 2$, $y = 1$. In general, Eq. (2.8) can have from 0 to d simple solutions. In particular, it has d simple solutions if all solutions to (2.9) are different and rational.

In the general case, every solution (x_1, x_2, \dots, x_n) to (2.7) produces an infinite family of solutions of the form $(kx_1, kx_2, \dots, kx_n)$, $k \in \mathbb{Z}$, hence we will count only *simple* solutions, that is, with no common factor, and such that the first non-zero

component of (x_1, x_2, \dots, x_n) is positive. Note that the solution $(0, 0, \dots, 0)$ does not count as simple, because the common factor condition does not hold for it.

Counting Solutions to $x + y + z = 0$

The next case is $n = 3$, and even one of the simplest equations in three variables

$$x + y + z = 0, \quad (2.10)$$

has infinitely many simple solutions (x, y, z) , for example, $(1, 0, -1)$, $(1, 1, -2)$, $(1, 2, -3)$, $(1, 3, -4)$, and so on. The interesting question is, for each B , what is the number $N_F(B)$ of simple solutions to (2.7) such that $\max\{|x|, |y|, |z|\} \leq B$. For Eq. (2.10), $N_F(0) = 0$, $N_F(1) = 3$ (simple solutions are $(0, 1, -1)$, $(1, 0, -1)$, and $(1, -1, 0)$), $N_F(2) = 6$ (new simple solutions are $(1, -2, 1)$, $(1, 1, -2)$, and $(2, -1, -1)$), and so on. Geometrically, Eq. (2.10) describes a plane in the coordinate space, and the condition $\max\{|x|, |y|, |z|\} \leq B$ describes a cube, so we are interested in studying points with integer coordinates which belong to the intersection of the plane and the cube. This intersection is a hexagon, and it is convenient to look at its projection on the plane $z = 0$, see Fig. 2.6. Each marked point with coordinates (x_0, y_0) in the figure corresponds to a simple solution $(x, y, z) = (x_0, y_0, -x_0 - y_0)$ of (2.10) with $\max\{|x|, |y|, |z|\} \leq 10$.

How many such solutions are there for some large B ? Well, in any simple solution $x \geq 0$, hence there are at most $B + 1$ possible choices for integer x in the range $[0, B]$. For each x , there are at most $2B + 1$ possible choices for integer y in the range $[-B, B]$, and then $z = -x - y$ is determined uniquely. Hence, for the number $N_F(B)$ of simple solutions we have an easy estimate

$$N_F(B) \leq (B + 1)(2B + 1) \leq 6B^2. \quad (2.11)$$

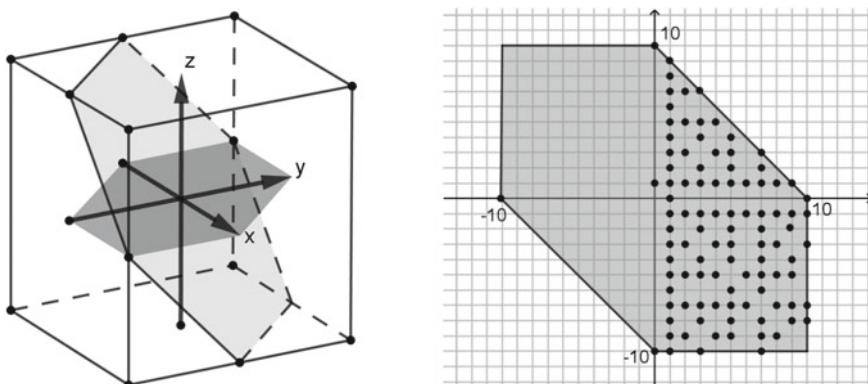


Fig. 2.6 Simple solutions to (2.10) for $B = 10$

The constant 6 here is not optimal, and can be improved very easily. But the true question is not about the constant but about the exponent 2. Can we have a better exponent, for example, an estimate of the form

$$N_F(B) \leq cB^{1.99},$$

for some constant c ? It turns out that we cannot. For example, it is known that at least $B/\ln B$ integers in the range $[1, B]$ are prime, and, for every prime x , at least $\frac{x-1}{x}B$ choices of y work (say, choose all $y \in [-B, -1]$ not divisible by x). Because $\frac{x-1}{x}B \geq B/2$, this results in at least $B^2/2\ln B$ simple solutions, and, for sufficiently large B ,

$$\frac{B^2}{2\ln B} > cB^{1.99}.$$

Homogeneous Equations in $n = 3$ Variables

For the general linear equation in three variables

$$ax + by + cz = 0,$$

where a, b, c are integer coefficients, $abc \neq 0$, the bound (2.11) remains correct, and a similar argument shows that the exponent 2 in it is the best possible.

Now, let us estimate $N_F(B)$ for a form $F(x, y, z)$ of arbitrary degree d . By a similar argument, there are at most $B + 1$ choices for x , at most $2B + 1$ for y , and, for every fixed x and y , z can be found from a polynomial equation of the form $a_0z^d + a_1z^{d-1} + \dots + a_{d-1}z + a_d = 0$. If not all coefficients are 0, there are at most d solutions, resulting in the estimate $N_F(B) \leq d(B + 1)(2B + 1) \leq 6dB^2$. Similar bounds can be proved for the case $a_0 = a_1 = \dots = a_{d-1} = a_d = 0$, resulting in the final estimate

$$N_F(B) \leq C_d B^2, \tag{2.12}$$

where C_d is a constant depending on d . In general, the exponent 2 in (2.12) cannot be improved, because the equation $(x + y + z)^d = 0$ has the same solutions as $x + y + z = 0$.

Irreducible Equations in $n = 3$ Variables

However, the equation $(x + y + z)^d = 0$ is somewhat artificial: it is *really* the linear equation $x + y + z = 0$ artificially written as an equation of degree d . To study forms F whose *actual* degree is d , we will assume that F is *irreducible*, that is, cannot be written as a product of polynomials of lower degree with rational coefficients. For such forms, Pila [306] showed in 1995 that we can have an estimate

$$N_F(B) \leq C(d, \varepsilon)B^{1+1/d+\varepsilon}, \quad (2.13)$$

for any $\varepsilon > 0$, where $C(d, \varepsilon)$ is a constant depending on d and ε . For example, for $d = 2$ we can choose any exponent greater than $1 + 1/d = 1.5$, which provides, for large B , a much better upper bound than $C_d \cdot B^2$. However, it was not clear whether this exponent is optimal or can be improved further. The following theorem of D.R. Heath-Brown [193] provides an optimal exponent.

Theorem 2.6 *For any irreducible form $F(x, y, z)$ of degree d in $n = 3$ variables, and any $\varepsilon > 0$,*

$$N_F(B) \leq C(d, \varepsilon)B^{2/d+\varepsilon}, \quad (2.14)$$

for some constant $C(d, \varepsilon)$ depending on d and ε .

Because $2/d < 1 + 1/d$ for $d > 1$, (2.14) is a substantial improvement over (2.13). Because equation $x^d - y^{d-1}z = 0$ has $B^{2/d}$ solutions of the form $(x, y, z) = (m^{d-1}n, m^d, n^d)$, $1 \leq n \leq \sqrt[d]{B}$, $1 \leq m \leq \sqrt[d]{B}$, (and a significant part of them are simple), the bound (2.14) is the best possible. Heath-Brown also proved generalizations of (2.14) in several directions, including some interesting estimates for the number of simple solutions to (2.7) in the next case, $n = 4$.

Geometrically, all real solutions to (2.7) form a curve if $n = 2$, and a surface if $n = 3$. An important question in mathematics is to understand how “dense” the rational points on a curve or surface are, where by “rational points” we mean rational solutions to (2.7). Because every rational solution can be transformed into an integer solution by multiplying by the common denominator, Theorem 2.6 directly contributes to this important research direction.

Reference

D.R. Heath-Brown, The density of rational points on curves and surfaces, *Annals of Mathematics* **155**-2, (2002), 553–598.

2.7 The Regular-Stochastic Dichotomy for Quadratic Polynomials

Dynamical Systems

In 1687, Isaac Newton wrote down differential equations describing how any number of bodies move subject to gravitation. However, these equations turned out to be extremely difficult to solve even in case of three bodies. Starting from some initial positions, the bodies move in a reasonably predictable way. However, some other initial positions turned out to lead to extremely complicated trajectories.

A system of bodies moving under gravitation is an example of a *dynamical system*, that is, a system evolving in time according to some fixed rule. One of the simplest

dynamical systems consists of a function $f : \mathbb{R} \rightarrow \mathbb{R}$ and an initial point $x_0 \in \mathbb{R}$, and the “dynamics” is just an iterative application of f to x_0 :

$$x_1 = f(x_0), \quad x_2 = f(x_1) = f(f(x_0)), \quad \dots, \quad x_n = f(x_{n-1}), \quad \dots \quad (2.15)$$

The function $f(f(\dots f(x) \dots))$ with n “iterations” will be denoted $f_n(x)$. In this notation, $x_n = f_n(x_0)$.

Dynamical Systems from Linear Functions

For a linear function, $f(x) = ax + b$, the sequence (2.15) is easy to analyse. First of all, if $a \neq 1$, there is a unique real number x^* such that $f(x^*) = x^*$. In fact, from the equation $ax^* + b = x^*$ we immediately see that $x^* = \frac{b}{1-a}$. If we start the sequence of iterations (2.15) with $x_0 = x^*$, then $x_1 = f(x_0) = f(x^*) = x^*$, $x_2 = f(x_1) = f(x^*) = x^*$, and so on: the whole sequence (2.15) becomes x^*, x^*, x^*, \dots . In general, a point x^* such that $f(x^*) = x^*$ is called a *fixed point* of a dynamical system (2.15).

It is convenient to perform the change of variables $y = x - x^*$, which puts the fixed point at the origin. Then $y_n = x_n - x^*$, $n = 0, 1, 2, \dots$, and

$$y_n = x_n - x^* = ax_{n-1} + b - x^* = a(y_{n-1} + x^*) + b - x^* = ay_{n-1} + (a-1)x^* + b = ay_{n-1}.$$

In other words, y_n is the sequence (2.15) corresponding to the function $g(y) = ay$. This immediately implies the general formula

$$y_n = a^n y_0, \quad n = 0, 1, 2, 3, \dots \quad (2.16)$$

If $|a| < 1$, then (2.16) implies that the sequence y_n converges to 0, which is a fixed point of g . In general, a fixed point x^* is called *attracting* if the sequence (2.15) converges to x^* for any x_0 belonging to the interval $(x^* - \varepsilon, x^* + \varepsilon)$ for some $\varepsilon > 0$.

If $|a| > 1$, then the fixed point $y^* = 0$ of g is not attracting. If we start with $y_0 = 0$, then $y_n = 0$ for all n . However, if $y_0 = \varepsilon$ for any $\varepsilon \neq 0$, however tiny, then $y_n = a^n \varepsilon$. If n increases, a^n quickly becomes huge, and y_n diverges to infinity.

The remaining case is $|a| = 1$. If $a = -1$, then $y_1 = -y_0$, $y_2 = -y_1 = y_0$, $y_3 = -y_2 = -y_0$, and the whole sequence becomes

$$y_0, -y_0, y_0, -y_0, y_0, -y_0, \dots$$

In general, if $x_k = x_0$ in (2.15) for some k , then $x_{k+1} = f(x_k) = f(x_0) = x_1$, $x_{k+2} = x_2$, and so on, and the whole sequence becomes

$$x_0, x_1, \dots, x_{k-1}, x_0, x_1, \dots, x_{k-1}, x_0, x_1, \dots, x_{k-1}, \dots$$

In this case, we call $(x_0, x_1, \dots, x_{k-1})$ a *cycle*.

Finally, if $a = 1$, our linear function is $f(x) = x + b$, and the sequence (2.15) is $x_n = x_0 + nb$. If $b = 0$, then $x_n = x_0$ for all n , that is, all points are fixed points. If $b \neq 0$, the system has no fixed points, and the sequence $x_n = x_0 + nb$ diverges for all x_0 .

Conjugate Functions

We have simplified the analysis of linear function dynamics using the change of variables $y = x - x^*$, or $x = y + x^* = y + \frac{b}{1-a}$. In general, if in (2.15) we make the “change of variables” $x = h(y)$ for some invertible function h , then $x_n = h(y_n)$, $n = 0, 1, 2, \dots$, and (2.15) reduces to

$$y_n = h^{-1}(x_n) = h^{-1}(f(x_{n-1})) = h^{-1}(f(h(y_{n-1}))), \quad n = 0, 1, 2, \dots$$

hence y_n is the sequence (2.15) corresponding to a function $g(y) = h^{-1}(f(h(y)))$. Such a function g is called *conjugate* to f . In our example, we have proved that the function $f(x) = ax + b$ is conjugate to a simpler function $g(y) = ay$. In general, it is always useful to find a change of variables h such that the conjugate function g is simpler than the initial function f . We will use this trick again in a later section.

Attracting Cycles and Regular Quadratic Functions

Having finished our analysis of the linear case, the next case is when f is a quadratic polynomial, and let us start with $f(x) = x^2$. If $x_0 = 0$, then $x_n = 0, \forall n$, so 0 is a fixed point of this dynamical system. For $x_0 = 0.5$, we have

$$x_1 = 0.25, \quad x_2 = 0.0625, \quad x_3 \approx 0.0039, \quad \dots, \quad x_n = (0.5)^{2^n}, \quad \dots$$

so the system quickly converges to the fixed point 0. The same happens for any x_0 such that $|x_0| < 1$, hence the fixed point $x^* = 0$ is attracting. Note that $x^* = 1$ is another fixed point of $f(x) = x^2$, but it is not attracting: for any x_0 even slightly smaller than 1 the sequence converges to 0, and for $x_0 > 1$ it diverges to infinity.

For $f(x) = x^2 - 1$, the behaviour of the sequence (2.15) may be more complicated. With $x_0 = 0$, we get

$$x_1 = 0^2 - 1 = -1, \quad x_2 = (-1)^2 - 1 = 0, \quad x_3 = -1, \dots, \quad x_{2n} = 0, \quad x_{2n-1} = -1, \dots$$

hence the system oscillates between 0 and -1 , or, in other words, has a cycle $(0, -1)$. For $x_0 = 0.5$, the sequence (2.15) becomes

$$0.5, -0.75, -0.4375, -0.8086, -0.3462, -0.8802, -0.2253, -0.9492, -0.0989, \dots \quad (2.17)$$

where only the first 4 digits of each x_n is given. We can see that the sequence becomes closer and closer to the cycle $(0, -1)$. A cycle $(x_0 = x^*, x_1, \dots, x_{k-1})$ is called *attracting* if there exists an $\varepsilon > 0$ such that the sequence (2.15) converges to this cycle for any x_0 such that $|x_0 - x^*| < \varepsilon$. A (quadratic) function f is called *regular* if the corresponding dynamical system (2.15) has an attracting cycle. For regular f , the sequence (2.15) is relatively easy to analyze, and for a typical initial point x_0 , we can (approximately) predict its value after sufficiently many iterates. For example, $x_{10,000}$ in (2.17) is approximately equal to 0.

The “Unpredictable” Dynamics of $f(x) = x^2 - 2$

However, an attracting cycle may not exist. For the function $f(x) = x^2 - 2$ and $x_0 = 0.5$, the sequence (2.15) starts with

$$0.5, -1.75, 1.0625, -0.8711, -1.2412, -0.4594, -1.7889, 1.2002, -0.5594, \dots \quad (2.18)$$

You can continue it as much as you want but will not notice any regular behaviour. Well, we always have $|x_n| < 2$, because $|f(x)| = |x^2 - 2| < 2$ for any x such that $|x| < 2$, but the numbers x_n seems to “walk” over the interval $(-2, 2)$ in a random way. In particular, it is not easy to tell what the 10^{1000} th term of (2.18) looks like, even approximately. The “unpredictable” behaviour of $f_n(x)$ for $f(x) = x^2 - 2$ is illustrated in Fig. 2.7.

Stochastic Quadratic Functions

For a complex-looking dynamical system, people at least try to answer questions like “What is the probability that the system will look this way in the future?” In our case, let us fix some large N , say, $N = 10^{1000}$, select an integer n such that $1 \leq n \leq N$ uniformly at random, and try to (approximately) find the probability $\mathbb{P}[a < x_n < b]$ for any a, b such that $-2 \leq a < b \leq 2$. For the sequence (2.18), the answer is

$$\mathbb{P}[a < x_n < b] \approx \int_a^b \frac{dx}{\pi \sqrt{4 - x^2}} = \frac{1}{\pi} (\arccos(a/2) - \arccos(b/2)).$$

Indeed, because $x_0 \in (-2, 2)$, there exists an angle α such that $x_0 = 2 \cos \alpha$. Then $x_1 = x_0^2 - 2 = (2 \cos \alpha)^2 - 2 = 2(2 \cos^2 \alpha - 1) = 2 \cos(2\alpha)$. Similarly, $x_2 = x_1^2 - 2 = 2(2 \cos^2 2\alpha - 1) = 2 \cos(4\alpha)$, and so on, $x_n = 2 \cos(2^n \alpha)$. (In other words, the function $f(x) = x^2 - 2$ is conjugate to a much simpler function $g(\alpha) = 2\alpha$ via the change of variables $x = 2 \cos \alpha$). Hence,

$$\mathbb{P}[a < x_n < b] = \mathbb{P}[a/2 < \cos(2^n \alpha) < b/2] \approx \frac{1}{\pi} (\arccos(a/2) - \arccos(b/2)),$$

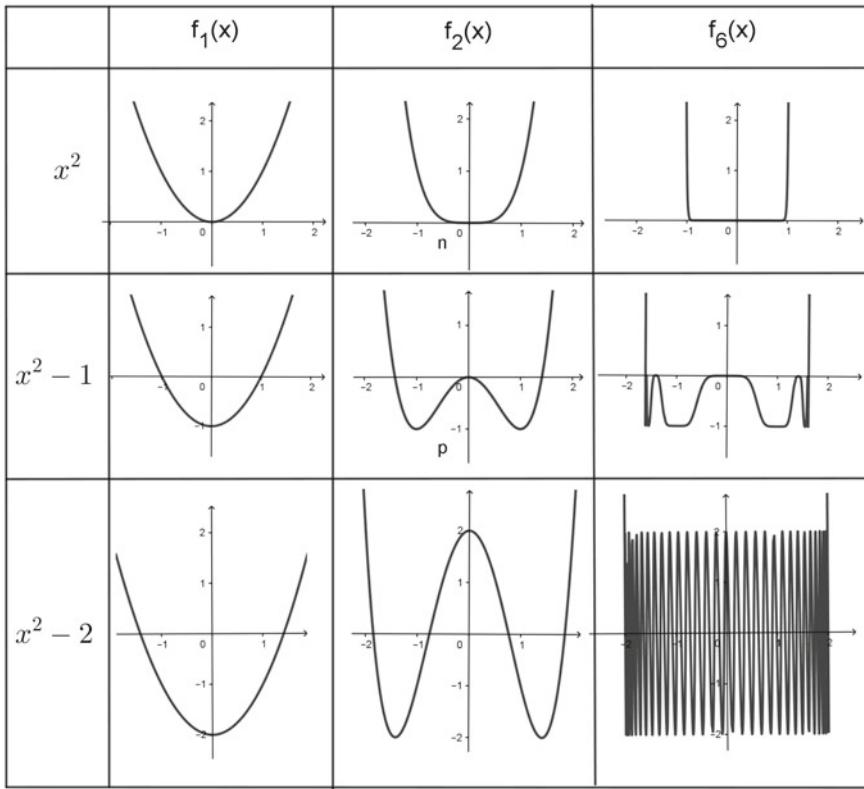


Fig. 2.7 Graphs of $f(x)$, $f_2(x) = f(f(x))$, and $f_6(x) = f(f(\dots(f(x))\dots))$ for $f(x) = x^2$, $f(x) = x^2 - 1$, and $f(x) = x^2 - 2$

where the approximate equality is based on the fact that $2^n\alpha$ is a random angle. If there exists a function ϕ such that

$$\lim_{N \rightarrow \infty} \mathbb{P}[a < x_n < b] = \int_a^b \phi(x) dx,$$

then the function f in (2.15) is called *stochastic*. Unlike in our example with $f(x) = x^2 - 2$, the function ϕ is usually difficult to compute analytically, but, if it can be estimated at least numerically, we have a satisfactory description of what to expect from the system in the long run.

The Regular-Stochastic Dichotomy

The main goal of studying dynamical systems is to describe the behaviour of “almost all” systems under “almost all” initial conditions, where the meaning of the phrase “almost all” is explained below. The following theorem of Michail Lyubich [256] achieves this goal for the simplest interesting dynamical systems, namely, it characterizes “almost all” systems of the form $f(x) = x^2 + c$, $c \in \mathbb{R}$.

Theorem 2.7 *Almost every real quadratic polynomial $f_c(x) = x^2 + c$, $c \in [-2, 1/4]$, is either regular or stochastic.*

The phrase “almost every” in Theorem 2.7 means “for all $c \in [-2, 1/4]$ except a set of measure 0”. A set of measure 0 is a set that can be covered by a collection of intervals of total length ε for any $\varepsilon > 0$. Equivalently, if we choose c from $[-2, 1/4]$ uniformly at random, Theorem 2.7 states that $x^2 + c$ will be either regular or stochastic with probability 1. For $c > 1/4$, $f(x) = x^2 + c = (x - 1/2)^2 + x + c - 1/4 \geq x + (c - 1/4)$, hence $x_n \geq x_{n-1} + (c - 1/4)$, and the sequence (2.15) diverges to $+\infty$. The same can be proved for any $c < -2$ and almost every x_0 , hence $c \in [-2, 1/4]$ is the only interesting range of parameters.

Although the statement of Theorem 2.7 is formulated in purely real terms, the proof substantially uses complex numbers, that is, numbers of the form $a + bi$, where $a, b \in \mathbb{R}$, and $i = \sqrt{-1}$, see e.g. Sect. 1.7 for details. This is in agreement with the classical Painleve–Hadamard Principle: “Between two truths in the real domain, the easiest and shortest path quite often passes through the complex domain”. The proof is also based on ideas of renormalization coming from physics, which provide us with a powerful method of penetrating into the small-scale structure of dynamical systems.

When facing a difficult problem, mathematicians often try to identify “the simplest interesting case”, solve it, and then apply the same methods to more complicated cases. The machinery developed to prove Theorem 2.7 has already been successfully extended to analyse more complex dynamical systems. Still simpler than the movement of m bodies under gravitation, but we now seem to be one step closer to a satisfactory solution of this problem as well.

Reference

M. Lyubich, Almost every real quadratic map is either regular or stochastic, *Annals of Mathematics* **156**-1, (2002), 1–78.

2.8 A Finitely Presented Group with an NP-Complete Word Problem

Groups and Their Sets of Generators

A *group* is a set G together with an operation \cdot such that (i) $a \cdot b \in G$ for all $a, b \in G$; (ii) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all $a, b, c \in G$; (iii) there exists an $e \in G$ (called the identity element of G) such that $a \cdot e = e \cdot a = a$ for all $a \in G$; and (iv) for every

$a \in G$, there exists an element $a^{-1} \in G$ (called the *inverse* of a), such that $a \cdot a^{-1} = a^{-1} \cdot a = e$. For example, the set of integers \mathbb{Z} with addition operation “+” forms a group, with identity element 0 and inverse element to any $a \in \mathbb{Z}$ being $-a$. See Sect. 1.5 for details and more examples of groups.

Note that all elements $a \in \mathbb{Z}$ can be written using only the element 1, its inverse -1 , and the group operation $+$, for example, $5 = 1 + 1 + 1 + 1 + 1$, $-3 = (-1) + (-1) + (-1)$, and so on. Similarly, we can represent any $a \in \mathbb{Z}$ as a sum of elements 2 and 3 (and their inverses), in particular, $0 = 2 + (-2)$, $1 = 3 + (-2)$, $-7 = (-3) + (-2) + (-2)$, and so on. This is not true, for example, for elements 4 and 6 (and their inverses), using which only even numbers can be represented. In general, we say that a subset $S \subset G$ is a *set of generators* for a group G if every element of G can be written as a product (here, by “product” we mean “the result of the group operation”) of elements of S and their inverses. For example, the sets $\{1\}$ and $\{2, 3\}$ are sets of generators for the group \mathbb{Z} , while the set $\{4, 6\}$ is not a set of generators. Any group having at least one finite set of generators S is called *finitely generated*. The products of elements of S are called *words*.

The Group of Symmetries of a Square

Groups also naturally arise in geometry. An *isometry* of the Euclidean plane \mathbb{R}^2 is a distance-preserving transformation of the plane, that is, a map $m : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, such that for any points A and B in the plane, $d(A, B) = d(m(A), m(B))$, where $d(A, B)$ is the distance between A and B . For example, any rotation of the plane around a fixed center O by an angle α is an isometry, as well as a reflection of the plane with respect to any line. For any two isometries m and n , we can define the isometry $m \cdot n$ as their composition (that is, perform m and n one after another), and then the set of all isometries form a group with identity element e corresponding to the “isometry” mapping each point into itself.

For a square $ABCD$, let us study the group of plane isometries mapping this square into itself, which is usually denoted by \mathbb{D}_4 . For example, let r be a clockwise rotation by 90° around the center O of the square. Then $r(A) = B$, $r(B) = C$, $r(C) = D$, $r(D) = A$, and the square is moved into itself, hence $r \in \mathbb{D}_4$. Also, $f \in \mathbb{D}_4$, where f is the reflection of the plane with respect to the diagonal of the square going from the top left vertex to the bottom right vertex. In fact, the group \mathbb{D}_4 consists of 8 transformations, and all of them can be expressed as a product (composition) of r and f , see Fig. 2.8. In other words, $\{r, f\}$ is a set of generators of \mathbb{D}_4 .

However, the representation of $a \in \mathbb{D}_4$ as a product of r and f is not unique. For example, reflecting the plane with respect to BD twice we map each point back into itself, hence $f^2 := f \cdot f = e$. Using this relation we can “simplify expressions”, for example, $f^3 = (f^2)f = ef = f$, $f^9rf^4 = fre = fr$, etc. Also, clockwise rotation by 90° , performed 4 times, is the “rotation” by 360° , which is e , hence $r^4 = e$. Again, we can use this relation for simplifications in the form $r^9 = r^4r^4r = r$, $r^8fr^5 = efr = fr$, etc. Finally, one can check that $rfrf = e$. For example, point A after rotation becomes B , after reflection in the line BD remains B , after rotation moves

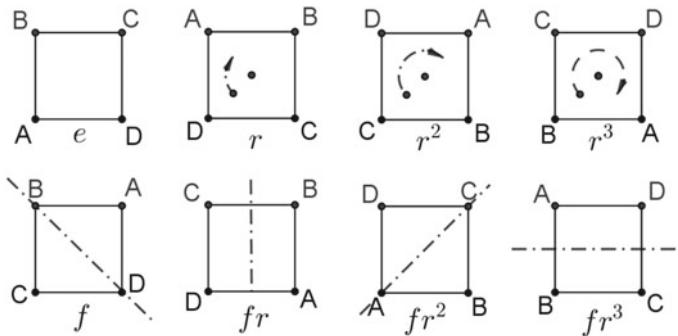


Fig. 2.8 Elements of the group of symmetries of a square

to C , and finally after reflection in BD moves back to A . It turns out that these three relations are sufficient to completely describe \mathbb{D}_4 : we can forget about geometry and define \mathbb{D}_4 as the group of all possible products of two elements r and f , subject to the constraints $r^4 = f^2 = rfrf = e$.

We say that a group is *finitely presented* if it can be described as a group of all possible products of elements from a finite set S (and their inverses), subject to a finite set of conditions (also called “relations”) in the form “some products are equal to the identity element”. For example, \mathbb{D}_4 is a finitely presented group with 2 generators and 3 relations. By definition, every finitely presented group is finitely generated, but the converse is not true.

The Word Problem

A fundamental problem about finitely generated groups is the *word problem*: given two words a and b , do they represent different elements? For example, the words f^3 and f trivially represent the same element in \mathbb{D}_4 , but what about the words r^3fr^2f and r ? It turns out that they are also the same, because

$$r^3fr^2f = r^3(rfrf)fr^2(rfrf)f = (r^4)fr(f^2)r^3fr(f^2) = fr^4fr = f^2r = r, \quad (2.19)$$

where we have repeatedly used the relations $r^4 = f^2 = rfrf = e$. However, it is not immediately obvious how one would guess to use the transformations shown in (2.19). Once they are written down, they are easy to *verify*, but how to *find* them?

The Subset Sum Problem and the Clique Problem

In mathematics, there are many problems which are difficult to solve but easy to verify once the solution is found. For example, given a set of integers a_1, a_2, \dots, a_n with even sum S , can we find a subset with sum $S/2$? In particular, let $n = 12$ and

$$a_1 = 285902, a_2 = 678203, a_3 = 186530, a_4 = 239856, a_5 = 239659, a_6 = 276591,$$

$$a_7 = 386032, a_8 = 907387, a_9 = 123974, a_{10} = 590254, a_{11} = 129865, a_{12} = 625125,$$

then $S = 4669378$, and $a_2 + a_8 + a_9 + a_{12} = 2334689 = S/2$. This is easy to verify, but how can one guess that we need to add the 2nd, 8th, 9th, and 12th numbers? Obviously, one could write a computer program which tries all $2^{12} = 4096$ subsets. However, if we had $n = 1000$ numbers instead of 12, checking all 2^n subsets would take even the fastest computers many millions of years to complete. This problem is called the *subset sum problem*.

As another example, let H be a graph (that is, a set of n vertices, some of them connected by edges), and consider whether there exist k vertices all connected to each other. This problem is called the *clique problem*. Once k such vertices are found, it is easy to verify that they are indeed all pairwise connected. For example, if $n = 1000$ and $k = 100$, and someone marked for you 100 vertices and claimed that they solve the problem, all you need to do is to look at all pairs of marked vertices and check each pair is indeed connected in H . Because there are less than $k^2 = 10,000$ such pairs, a computer can verify this for you in a second.

However, if $k = 100$ pairwise connected vertices are not marked, does there exist an efficient method to *find* them? Trying all possible subsets is not an option, because the number of 100-vertex subsets in a 1000-vertex graph is much higher than the number of atoms in the observable universe.

Decision Problems, Algorithms, and the Class P

In general, a *decision problem* is any yes-or-no question on an infinite set of inputs. An *algorithm* for a decision problem is a computer program which takes any input I of length n , works for some time $T(I)$, and then outputs the correct result. An algorithm is called *polynomial* if there is a polynomial Q such that $T(I) \leq Q(n)$ for all inputs I . The exact form of Q depends on the speed of the computer, and also how exactly we measure the time (in seconds, minutes, etc.), but the definition of a polynomial algorithm is universal: if algorithm A is polynomial on some computer, it is so on any other computer, just with a different polynomial Q . The class of all decision problems having a polynomial algorithm is called P .

For example, a *triangle* in a graph is a set of three vertices A, B, C , which are all connected by edges. The problem “given a graph G , check if it contains a triangle” can be solved by checking for all triples of vertices whether they form a triangle or not. If the graph has n vertices, there are less than n^3 triples of vertices, hence the whole algorithm will run in time at most Cn^3 , where C is a constant which depends on the speed of your computer. Hence, this problem belongs to the class P .

Verification of Solution, and the Class NP

Now assume that, in a decision problem, (i) for every instance where the answer is “yes”, there exists a proof of the fact that the answer is indeed “yes”, and (ii) there exists a polynomial algorithm which can verify this proof, that is, outputs “yes” if the proof is correct, and “no” otherwise. The class of all such decision problems is called NP . For example, in the subset sum problem, if the answer is “yes”, and there is a subset with sum $S/2$, then (i) we can just list the elements of this subset as a proof, and (ii) we can easily and quickly verify this proof, by just adding up the listed elements, and checking that the sum is indeed $S/2$. Hence, this problem belongs to class NP . And the same is true for the clique problem.

Obviously, every problem belonging to class P also belongs to NP , but the converse is an open question. In particular, while the subset sum problem and the clique problem are in NP , we do not know if there exists a polynomial algorithm for solving any of these problems, that is, we do not know if they are in P or not. Amazingly, if we can find a polynomial algorithm for one of these problems, then we automatically find a polynomial algorithm for the other one! In fact, a polynomial algorithm for any of these problems would imply that $P = NP$, that is, every problem whose solution can be *verified* in polynomial time could also be *solved* in polynomial time. The question “Is P equal to NP ?” is one of the seven Millennium Prize problems [85], for which a million dollar prize is promised for a correct solution.

NP-Complete Problems

A problem in NP such that a polynomial algorithm for it would imply that $P = NP$ is called NP -complete. The subset sum problem and the clique problem are examples of NP -complete problems, and there are thousands of others. Developing a fast algorithm for *any* of them would solve them *all* at once, and bring you a million dollars. On the other hand, most people believe that in fact $P \neq NP$, hence proving that any problem is NP -complete is a strong indication that it just cannot be efficiently solved. The following theorem of Mark V. Sapir, Jean-Camille Birget and Eliyahu Rips [332] states that the word problem for a finitely presented group may be NP -complete.

Theorem 2.8 *There exists a finitely presented group for which the word problem is NP -complete.*

Theorem 2.8 states that there exists a finitely presented group G in which, for any two words representing the same element, there exists an efficiently verifiable proof, like (2.19), that they are indeed the same. However, an efficient algorithm for *finding* such a proof does not exist, unless $P = NP$.

Reference

M. Sapir, J.-C. Birget and E. Rips, Isoperimetric and isodiametric functions of groups, *Annals of Mathematics* **156**-2, (2002), 345–466.

2.9 Finitely Generated Groups with a Word Problem in NP

This section uses the definitions introduced in Sect. 2.8, and continues the discussion started there.

Finding and Checking the Solution of a Word Problem

Theorem 2.8, described in Sect. 2.8, implies that there exists a finitely presented group G^* whose word problem is in NP (which means that for any two words, representing the same element, there exists an efficiently verifiable proof that they are indeed the same), but it is difficult to *find* such a proof. However, the word problem for the group G^* is not the most difficult one: there are finitely presented groups for which the word problem is not even in NP .

For example, consider the following problem: for a given graph H and positive integer k , it is true that among any k vertices in H we can find two which are *not* connected? In this problem, if the answer is “No”, this is easy to prove: any set of k pairwise connected vertices provides a counter-example. However, if the answer is “Yes”, there is no easy way to verify this. In fact, this problem is just the NP -complete clique problem discussed in Sect. 2.8, with the answers “Yes” and “No” interchanged. Such problems are called *coNP-complete*. There is a finitely presented group G with *coNP*-complete word problem! That is, if two words a and b in G represent *different* elements, there is a short proof of this fact. However, if they represent the same element, no such proof is known to exist.

At first glance, this looks counter-intuitive. If two words a and b in any finitely presented group G represent the same element, then we can always prove this by writing down a chain of transformation like (2.19). The problem is that, for some groups, this chain can be very long, and no polynomial algorithm can check it, because it has no time even to read it!

Isoperimetric Functions

In (2.19), we first insert $e = rfrf$ in the formula (twice), then delete $e = r^4$, then $e = f^2$ (twice), then again $e = r^4$, and finally again $e = f^2$, totalling 7 operations (2 insertions and 5 deletions). Actually, checking that the words r^3fr^2f and r in (2.19) are the same is equivalent to checking that the word $a = r^3fr^2fr^{-1}$ represents the identity element. For any word a representing the identity element e , let $T(a)$ be the minimal number of operations (insertions and deletions) required to transform a to e . A function $f : \mathbb{N} \rightarrow \mathbb{N}$ is called an *isoperimetric function* of G if

$$T(a) \leq f(|a|) \tag{2.20}$$

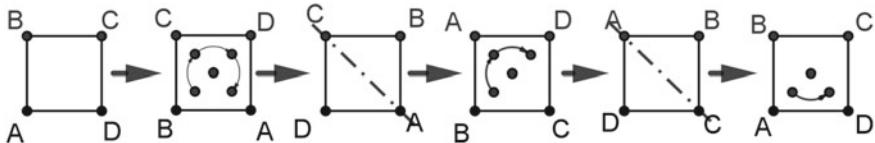


Fig. 2.9 Illustration of $r^3 fr^2 fr^{-1} = e$

for every a representing e , where $|a|$ is the length of the word a . For example, in (2.19) we need 7 operations for a word $a = r^3 fr^2 fr^{-1}$ of length 8, and one can check that this is the best possible for this word, hence $7 \leq f(8)$ should hold for any isoperimetric function f of the group \mathbb{D}_4 .

Assume that a finitely presented group G has a polynomial isoperimetric function, for example, $f(n) = n^3$. Then, for every word a representing e , there exists a chain like (2.19) transforming a into e in at most $|a|^3$ operations. This chain can be easily verified by a computer program, hence this proves that G is in NP . A similar argument works for any finitely generated subgroup H of G (a subset H of a group G is called a *subgroup* if it is also forms a group with the same operation), resulting in the following statement.

- (*) If a finitely generated group H is a subgroup of a finitely presented group G with a polynomial isoperimetric function, then the word problem of H is in NP .

Alternative Approaches to the Word Problem

Polynomial isoperimetric functions do not always exist. In particular, one can find a finitely presented group G which has infinitely many words a representing e such that the shortest chain of operations transforming a to e is longer than, say, $2^{|a|}$. How can we then prove that a indeed represents e ? Writing down the full chain is not an option, because it would take too long for any computer program to even read it, let alone verify it.

Sometimes, there are indeed better proofs. For example, to prove that $a = r^3 fr^2 fr^{-1}$ represents the identity element in \mathbb{D}_4 , we can, instead of using (2.19), directly check that a sends every vertex of the square to itself, see Fig. 2.9. This method seems to be much easier and more intuitive than (2.19).

A Criterion for the Word Problem Belonging to NP

This example leads us to the intuition that the word problem is in NP for a much wider class of groups than those described by (*). After all, “the word problem of H is in NP ” means that there *exists* an efficiently verifiable method for proving that two words represent the same element in H . This method can use any special properties of the group, in a similar way as we have used geometric intuition to

analyse words in \mathbb{D}_4 . In contrast, condition (*) describes the special case when a “chain of transformations” method like (2.19) works.

However, the following theorem of Birget, Ol'shanskii, Rips and Sapir [56] states that this intuition is wrong, and, in fact, no other groups except those satisfying (*) have the word problem in NP .

Theorem 2.9 *The word problem of a finitely generated group H is in NP if and only if H is a subgroup of a finitely presented group G with a polynomial isoperimetric function.*

Theorem 2.9 states that if, for a finitely generated group H , we have developed a clever efficient method to prove that the words are the same, then, in fact, the word equivalence could be demonstrated using a “chain of transformations” method as in (2.19). We just need to apply this method not directly to H , but to a larger group G , described in Theorem 2.9.

Groups with an Undecidable Word Problem

There are finitely generated, and even finitely presented groups, for which the word problem is not in NP , that is, there is no efficient way to prove that two words represent the same element. In fact, there are groups for which the word problem cannot be solved by *any* algorithm, not necessary polynomial. Such problems are called *undecidable*. In 1986 Collins [99] constructed an explicit finitely presented group H^* with 10 generators and 27 relations having an undecidable word problem. If, for any word a of length n representing e in H^* , there were a proof that this length is bounded by some fixed polynomial $P(n)$, we could check if $a = e$ just by trying all possible proofs of length up to $P(n)$. If some proof works, then $a = e$, otherwise $a \neq e$. This would require a very large, but still finite amount of time. Collins proved that there is no such algorithm, and therefore the word problem of H^* is not in NP . Moreover, the same argument works if we replace the polynomial $P(n)$ by a function 2^n , 2^{2^n} , etc. In particular, this implies that there are infinitely many words in H^* of length n which can be transformed to the identity element by transformations like in (2.19), but more than 2^{2^n} operations are required for this. And the same statement remains correct if we replace 2^{2^n} by *any* formula one can explicitly write down!

However, examples of groups like H^* are rather artificial. The beauty and importance of Theorem 2.9 is that for many finitely generated groups H naturally arising in applications, people have already developed efficient methods for proving word equivalence, hence we know that their word problems are in NP . We can now apply Theorem 2.9 to conclude that all such groups are in fact subgroups of a finitely presented group G with a polynomial isoperimetric function. Because finitely presented groups (and their subgroups) are better understood, this gives us a lot of useful information about the original group H .

Reference

J.-C. Birget, A.Yu. Ol'shanskii, E. Rips and M.V. Sapir, Isoperimetric functions of groups and computational complexity of the word problem, *Annals of Mathematics* **156**-2, (2002), 467–518.

2.10 Positive Noncommutative Polynomials are Sums of Squares

The “Sum of Squares” Method for Proving Inequalities

How would you prove that

$$P(x, y, z) := x^2 + y^2 + z^2 - xy - yz - zx \geq 0, \quad \forall x, y, z \in \mathbb{R} \quad (2.21)$$

Perhaps the most straightforward approach is to observe that

$$P(x, y, z) = \left(\frac{x-y}{\sqrt{2}} \right)^2 + \left(\frac{y-z}{\sqrt{2}} \right)^2 + \left(\frac{z-x}{\sqrt{2}} \right)^2,$$

which, being a sum of squares, is always non-negative. This method is known as *sum of squares* (SoS) method of proving inequalities. There exist efficient computer algorithms which can find representations of a polynomial as a sum of squares (provided that it exists), and, in this way, apply the SoS method to prove a broad range of inequalities with polynomials.

Unfortunately, the SoS method does not always work. For example, the inequality

$$1 - 3x^2y^2 + x^2y^4 + x^4y^2 \geq 0$$

is true for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$, see Fig. 2.10 for a graphical illustration, but the polynomial on the left-hand side cannot be expressed as a sum of squares of other polynomials with real coefficients. This example was found by Motzkin [284] in 1967.

Matrices, Their Sums, Products, and Transpositions

The SoS method can be extended to prove inequalities involving matrices, for example, to prove that the matrix polynomial

$$Q(A, A^T, B, B^T) := A^T A + 2A^T B + 2B^T A + 5B^T B \quad (2.22)$$

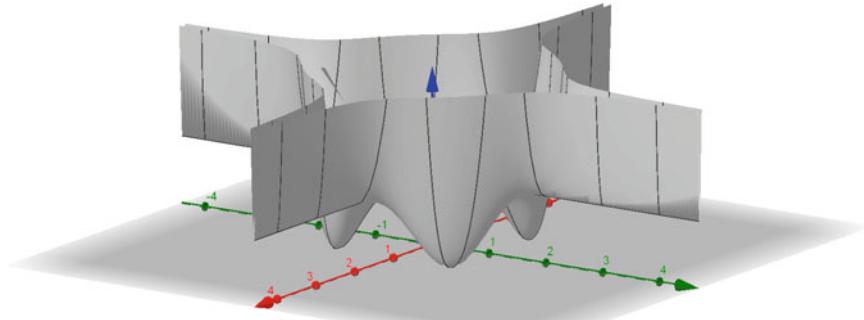


Fig. 2.10 The graph of $P(x, y) = 1 - 3x^2y^2 + x^2y^4 + x^4y^2$

always returns a “non-negative” matrix for any $n \times n$ matrices A and B as input (and any natural n). To understand (2.22), we need some definitions.

An $n \times m$ matrix is a 2-dimensional array of elements of the form

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix}.$$

$n \times 1$ matrices are called *vectors*. The sum of two $n \times m$ matrices A and B is an $n \times m$ matrix $C = A + B$ with elements $c_{ij} = a_{ij} + b_{ij}$, $\forall i, j$. The product of an $n \times m$ matrix A and an $m \times l$ matrix B is an $n \times l$ matrix $C = AB$ with elements $c_{ij} = \sum_{k=1}^m a_{ik}b_{kj}$, $\forall i, j$, see Sect. 2.2.

For any $n \times m$ matrix A , its *transpose*, denoted A^T , is an $m \times n$ matrix with elements $a'_{ij} = a_{ji}$, that is, with columns and rows exchanged, i.e.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^T = \begin{bmatrix} a & c \\ b & d \end{bmatrix}, \quad \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}^T = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}, \quad \text{etc.}$$

It is easy to check that $(A + B)^T = A^T + B^T$ and $(AB)^T = B^T A^T$. An $n \times n$ matrix A such that $a_{ij} = a_{ji}$, $\forall i, j$, or, equivalently, $A^T = A$, is called *symmetric*.

Positive Semi-definite Matrices

It is not very clear what is meant by a “non-negative” matrix in (2.22), because non-negativity is defined only for numbers. There is a “standard” way to transform an $n \times n$ matrix to a number: We can multiply it by any $n \times 1$ vector x from the right

and by its transpose x^T from the left. The result is always a 1×1 matrix, that is, a number, which can then be checked for non-negativity. Formally, a symmetric $n \times n$ matrix A is called *positive semi-definite* if $x^T A x \geq 0$, $\forall x \in \mathbb{R}^n$.

For example,

$$\begin{bmatrix} x & y & z \end{bmatrix} \cdot \begin{bmatrix} 1 & -0.5 & -0.5 \\ -0.5 & 1 & -0.5 \\ -0.5 & -0.5 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = x^2 + y^2 + z^2 - xy - yz - zx,$$

and, by proving inequality (2.21), we show that the matrix $\begin{bmatrix} 1 & -0.5 & -0.5 \\ -0.5 & 1 & -0.5 \\ -0.5 & -0.5 & 1 \end{bmatrix}$ is positive semi-definite.

Matrix-Positive Polynomials

A polynomial $Q = Q(X_1, X_2, \dots, X_n, X_1^T, X_2^T, \dots, X_n^T)$ returning a positive semidefinite matrix for any $n \times n$ input matrices X_1, X_2, \dots, X_n (and any natural n) is called *matrix-positive*. Now, how can we use the SoS method to prove that the polynomial in (2.22) is matrix positive? Unfortunately, the “standard” matrix square $C^2 = C \cdot C$ is not even guaranteed to be symmetric (check!), let alone positive semi-definite. The correct analogues for squares in the SoS method are expressions of the form $C^T \cdot C$ for an $n \times n$ matrix C . This is always symmetric, and, for any vector $x \in \mathbb{R}^n$,

$$x^T (C^T C)x = (x^T C^T)Cx = (Cx)^T (Cx).$$

Here Cx is an $n \times 1$ vector with elements y_1, y_2, \dots, y_n , hence $(Cx)^T (Cx)$ is a number equal to $y_1^2 + y_2^2 + \dots + y_n^2 \geq 0$. Hence, $C^T \cdot C$ is always positive semi-definite. By a similar argument, any sum of the form

$$\sum_{i=1}^k C_i^T \cdot C_i \tag{2.23}$$

is positive semi-definite for any matrices C_1, C_2, \dots, C_n . Hence, to prove that the polynomial (2.22) is matrix-positive, it is sufficient to present it in the “sum of squares” form (2.23). In this case, this is easy:

$$Q(A, A^T, B, B^T) = A^T A + 2A^T B + 2B^T A + 5B^T B = (A + 2B)^T (A + 2B) + B^T B,$$

and the problem is solved.

Symmetric Matrix-Positive Polynomials are Sums of Squares

Now, the basic question about every method is how general is it, how often does it work? Obviously, in general, polynomials like $Q(A) = A \cdot A$ do not return a symmetric matrix, and hence cannot be represented in the form (2.23). We say that a polynomial Q is *symmetric* if $Q^T = Q$ for any “input” matrices. In particular,

$$Q^T = (A^T A + 2A^T B + 2B^T A + 5B^T B)^T = A^T A + 2B^T A + 2A^T B + 5B^T B = Q,$$

hence the polynomial Q in (2.22) is symmetric. The following theorem of Helton [197] states that the matrix SoS method (as described above) *always* works for symmetric matrix positive polynomials!

Theorem 2.10 *Any symmetric matrix positive polynomial*

$$Q = Q(X_1, X_2, \dots, X_n, X_1^T, X_2^T, \dots, X_n^T)$$

can be represented in the SoS form (2.23).

The representation (2.23), if it exists, can be found quite efficiently using a computer. Hence, by Theorem 2.10, a computer, using the SoS method, can efficiently answer all questions of the form “Is the given symmetric polynomial matrix-positive or not”?

Reference

J.W. Helton, “Positive” noncommutative polynomials are sums of squares, *Annals of Mathematics* **156**-2, (2002), 675–694.

2.11 The Only Space Isomorphic to Each of its Subspaces

Measuring the Diagonal of a 4-Dimensional Cube

Imagine 2-dimensional “animals” living inside a plane. For them, the plane is the whole universe, and they could not imagine anything outside it. They can easily draw a unit square and calculate (for example) the length of its diagonal, which is $\sqrt{2}$, but they could hardly imagine a 3-dimensional cube.

Perhaps the usual 3-dimensional space we live in is in fact part of a larger, 4-dimensional space. How can we describe a space we cannot see, and, for example, calculate the diagonal of a 4-dimensional cube? For this, it is convenient to use coordinates and vectors. In the plane, we can fix any point O to be the *origin*, and then every point A can be associated with the vector \mathbf{OA} , which can be described using two coordinates (x, y) . We can then add vectors $((x_1, y_1) + (x_2, y_2)) = (x_1 + x_2, y_1 + y_2)$, multiply by a constant $(\lambda \cdot (x, y) = (\lambda x, \lambda y))$, and calculate the length

of every vector using the formula $|(\mathbf{x}, \mathbf{y})| = \sqrt{x^2 + y^2}$. In particular, the unit square can have vertices with coordinates $(0, 0)$, $(0, 1)$, $(1, 1)$, and $(1, 0)$, and the length of the diagonal is given by $|(1, 1)| = \sqrt{1^2 + 1^2} = \sqrt{2}$, as expected.

Similarly, 3-dimensional space can be described as a set of triples (x, y, z) . With this approach, it is not hard to go ahead and describe 4-dimensional space as a set of quadruples (x, y, z, t) , with length given by $|(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{t})| = \sqrt{x^2 + y^2 + z^2 + t^2}$. The 4-dimensional unit cube is an object with 16 vertices with coordinates (x_1, x_2, x_3, x_4) , where each x_i is either 0 or 1. The length of the diagonal is $|(1, 1, 1, 1)| = \sqrt{1^2 + 1^2 + 1^2 + 1^2} = 2$, see Fig. 2.11.

Infinite Dimensions and l^2 -Space

Who knows, maybe our 3-dimensional space is actually part of an even larger, 5-dimensional, or 100-dimensional space. For every finite n , n -dimensional space can be described as a set of vectors (x_1, x_2, \dots, x_n) with length $\sqrt{\sum_{i=1}^n x_i^2}$. Can we go further and imagine a space with *infinite* dimension? In such a space, every vector could have an infinite number of coordinates and could be written as $\mathbf{x} = (x_1, x_2, \dots, x_n, \dots)$. There is no problem in defining addition $(x_1, x_2, \dots, x_n, \dots) + (y_1, y_2, \dots, y_n, \dots) = (x_1 + y_1, x_2 + y_2, \dots, x_n + y_n, \dots)$ and multiplication by a constant $\lambda \cdot \mathbf{x} = (\lambda x_1, \lambda x_2, \dots, \lambda x_n, \dots)$, but how do we calculate length? By analogy with the finite-dimensional cases we could write it as the square root of the sum of squares of all coordinates, that is,

$$|\mathbf{x}| = \sqrt{\sum_{i=1}^{\infty} x_i^2}, \quad (2.24)$$

but how do we understand the infinite sum in (2.24)? What is the sum of squares of all coordinates, say, for $\mathbf{x} = (1, 1/2, 1/4, \dots, 1/2^n, \dots)$? Calculations shows that $1^2 = 1$, $1^2 + (1/2)^2 = 1.25$, $1^2 + (1/2)^2 + (1/4)^2 = 1.3125$, $1^2 + (1/2)^2 + (1/4)^2 + (1/8)^2 \approx 1.328$, Adding more and more terms takes us closer and

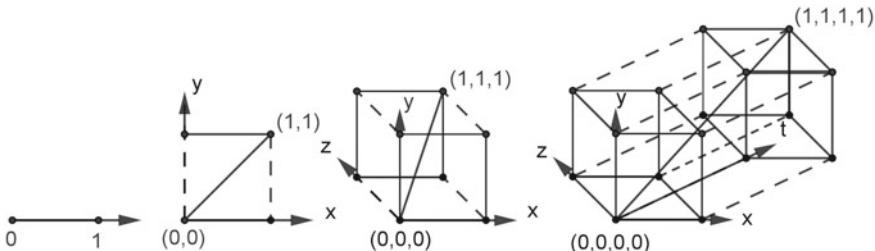


Fig. 2.11 Line segment, square, cube, and four-dimensional cube

closer to $\frac{4}{3}$. Formally, however small we choose $\varepsilon > 0$, for a sufficiently large number of terms n we get $|\frac{4}{3} - \sum_{i=1}^n x_i^2| < \varepsilon$. In this case, we say that $\frac{4}{3}$ is the *limit* of the sequence of sums, that is, $\lim_{n \rightarrow \infty} \sum_{i=1}^n x_i^2 = \frac{4}{3}$. In general, any infinite sum $\sum_{i=1}^{\infty} a_i$ is defined as the limit $\lim_{n \rightarrow \infty} \sum_{i=1}^n a_i$ of partial sums, if the latter exists. In particular, the length of our vector x is $|x| = \sqrt{\frac{4}{3}}$.

Obviously, the limit does not always exist. For example, for the “vector” $x = (1, 1, \dots, 1, \dots)$ the partial sums $\sum_{i=1}^n x_i^2 = n$, and (2.24) would give $|x| = \lim_{n \rightarrow \infty} n = \infty$. It is unintuitive to have vectors with infinite length, so let us just exclude them from consideration, and consider the set of vectors $x = (x_1, x_2, \dots, x_n, \dots)$ such that $|x| := \sqrt{\sum_{i=1}^{\infty} x_i^2} < \infty$. This is called l^2 -space.

Vector Spaces and Their Subspaces

In general, a *vector space* is any set V of objects (called *vectors*) for which we can define addition and multiplication by a constant, such that the usual laws ($x + y = y + x$, $x + (y + z) = (x + y) + z$, $\lambda(x + y) = \lambda x + \lambda y$, $(\lambda_1 + \lambda_2)x = \lambda_1 x + \lambda_2 x$, $\lambda_1(\lambda_2 x) = (\lambda_1 \lambda_2)x$, $1 \cdot x = x$) hold. There should also exist a *zero vector* $\mathbf{0}$ such that $x + \mathbf{0} = x$ for all $x \in V$. Finally, for any $x \in V$ there exists an *inverse* $-x \in V$ such that $x + (-x) = \mathbf{0}$. A vector space is called *normed* if we can associate with every $x \in V$ its length, or norm $\|x\|$, such that (i) $\|\mathbf{0}\| = 0$ but $\|x\| > 0$ if $x \neq \mathbf{0}$, (ii) $\|\lambda x\| = |\lambda| \|x\|$, and (iii) $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality). These properties hold in all the examples above, although verification of the triangle inequality in l^2 -space requires some work.

A subset X of a vector space V is called a *subspace* if $x + y \in X$ for every $x, y \in X$ and $\lambda x \in X$ for every $x \in X$, $\lambda \in \mathbb{R}$. A subspace X of a normed vector space is called *closed* if $x_n \in X$, $\forall n$ and $\lim_{n \rightarrow \infty} x_n = x$ (that is, for every $\varepsilon > 0$ we have $\|x - \sum_{i=1}^n x_i\| < \varepsilon$ for sufficiently large n) implies $x \in X$. For example, let V be the plane with vectors described by coordinates (x_1, x_2) , X_1 be the line given by the equation $x_1 = 0$ (that is, set of vectors in the form $(0, x_2)$), and X_2 be the line given by the equation $x_2 = 0$. Then X_1 and X_2 are both closed subspaces of V .

Isomorphic Vector Spaces

Normed vector spaces V and V' are called *isomorphic* if there exists a bijection $f : V \rightarrow V'$ such that $f(x + y) = f(x) + f(y)$, $\forall x, y \in V$, $f(\lambda x) = \lambda f(x)$, $\forall x \in V, \lambda \in \mathbb{R}$ and there exist constants $m > 0$ and $M > 0$ such that $m\|x\|_V \leq \|f(x)\|_{V'} \leq M\|x\|_V$.

$M\|x\|_V, \forall x \in V$. The idea is that V and V' are essentially the same space, just “rotated”. For example, the lines X_1 and X_2 defined above are isomorphic with $f(0, a) = (a, 0)$, $a \in \mathbb{R}$. However, the full 2-dimensional plane V is isomorphic to neither X_1 nor X_2 , nor any other line. Similarly, 3-dimensional space has many subspaces (planes, lines), but cannot be isomorphic to any of them.

Somewhat surprisingly, the situation is different for the infinite-dimensional space l^2 . Let X be the subspace of l^2 consisting of vectors of the form $x = (0, x_2, x_3, \dots, x_n, \dots)$, that is, the first coordinate is 0. Let $f : l^2 \rightarrow X$ be the function transforming any $x = (x_1, x_2, \dots, x_n, \dots)$ into $(0, x_1, x_2, \dots, x_n, \dots)$, that is, it adds 0 at the beginning and shifts all the coordinates. It is easy to check that f satisfies all the properties above, hence l^2 is isomorphic to its own subspace X ! Actually, by a similar argument, one can prove that l^2 is isomorphic to *every* infinite-dimensional closed subspace of itself.

A Unique Property of l^2 -Space

A normed vector space V is called a *Banach space* if for every sequence of vectors $x_1, x_2, \dots, x_n, \dots$ such that $\sum_{n=1}^{\infty} \|x_n\| < \infty$, there exists an $x \in V$ such that $\sum_{n=1}^{\infty} x_n = x$. One can check that the examples above, including l^2 -space, are Banach spaces, but this is just a tiny portion of the interesting Banach spaces studied in mathematics. For example, the space of continuous functions on $[0, 1]$ with norm $\|f\| = \int_0^1 |f(x)|dx$ is a Banach space as well.

However, despite the variety of examples of Banach spaces available, nobody could find another one which, similar to l^2 -space, is isomorphic to every infinite-dimensional closed subspace of itself. In fact, in 1932 Banach [34] asked whether l^2 -space is *the only* such example. After 70 years, this classical question was finally answered by the following theorem of Gowers [167], and the answer turned out to be “yes”.

Theorem 2.11 l^2 -space is (up to isomorphism) the only infinite-dimensional Banach space which is isomorphic to every infinite-dimensional closed subspace of itself.

As often happens, to answer an old classical question, the author needed to develop new tools which have many more applications. In the same paper, Gowers applies the developed methods to begin a “classification” of infinite-dimensional Banach spaces such that “knowing that a space belongs to a particular class gives a lot of information about the structure of the space”.

About the Proof: The Dichotomy Method and Ramsey Theorems

Let us say, briefly and informally, a few words about the proof of Theorem 2.11, which is based on the *dichotomy* method. In general, a dichotomy is just a partition of a set into two parts (subsets). For example, all children in a class can be partitioned

into two groups: boys and girls. In other words, *every child in the class is either a boy or a girl*. Hence, if we know that (a) all boys in the class like chocolate, and (b) all girls in the class like chocolate, it follows immediately that all children in the class like chocolate. In a similar way, it was known that (a) Theorem 2.11 is correct for all Banach spaces having some special property *A*, and (b) Theorem 2.11 is correct for all Banach spaces having some special property *B*. What Gowers proved is the dichotomy theorem stating that (*) *every Banach space has either property A or property B*.

Another example of a dichotomy is the statement “In every group of 3 people either there are 2 who know each other, or all 3 do not know each other”. In general, statements of the form “In every group of n people either there are k who all know each other, or there are m who all do not know each other” are known as Ramsey Theorems. Gowers proved a theorem of a somewhat similar nature, but for infinitely large “groups”, which he called “an infinite Ramsey Theorem”. He then deduced from it his dichotomy theorem (*), and Theorem 2.11 followed immediately.

Reference

W.T. Gowers, An infinite Ramsey theorem and some Banach-space dichotomies. *Annals of Mathematics* **156**-3, (2002), 797–833.

2.12 Counting Matrices with Some Special Properties

Alternating-Sign Matrices

A square matrix is an $n \times n$ array of elements of the form

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}.$$

It is called an *alternating-sign matrix* (ASM) if (i) all a_{ij} are equal to -1 , 0 , or 1 , (ii) the sum of elements in each row and column is 1 , and (iii) the non-zero elements in each row and column alternate in sign. The smallest examples of ASMs are

$$\begin{aligned} [1], \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 1 & -1 & 1 \\ 0 & 1 & 0 \end{bmatrix}. \end{aligned}$$

Such matrices have a close connection with several areas of mathematics, and even arise in the so-called “six-vertex model” for ice modelling in statistical mechanics.

How Many ASMs Are There?

The basic question is *how many* ASMs exist of a given size. From the list above, we can see that there is 1 ACM of size 1×1 , 2 ASMs of size 2×2 , and 7 ASMs of size 3×3 . Computer experiments show that there are 42 ASMs of size 4×4 , including, for example,

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

429 ASMs of size 5×5 , 7436 ASMs of size 6×6 , and so on. One might observe that these numbers are products of relatively small multipliers, for example, $7436 = 2 \cdot 2 \cdot 11 \cdot 13 \cdot 13$, which is a hint that there should be some formula for them involving products. Such a formula was discovered by David Robbins and Howard Rumsey in the early 1980s: the number of $n \times n$ ASMs should be

$$\frac{1! \cdot 4! \cdot 7! \cdots (3n-2)!}{n! \cdot (n+1)! \cdots (2n-1)!},$$

where $k!$ denotes the product $1 \cdot 2 \cdots k$. In particular, for $n=1$ we get $\frac{1!}{1!}=1$, for $n=2$: $\frac{1! \cdot 4!}{2! \cdot 3!} = \frac{24}{2 \cdot 6} = 2$, for $n=3$:

$$\frac{1! \cdot 4! \cdot 7!}{3! \cdot 4! \cdot 5!} = \frac{1}{3!} \frac{7!}{5!} = \frac{1}{6} (6 \times 7) = 7,$$

and so on. It took more than 10 years before Doron Zeilberger [407] proved that this formula indeed works in general. Later, Greg Kuperberg [233] found a substantially simpler proof of this result.

Vertically Symmetric ASMs

It is also important to count the number of ASMs obeying some special symmetries. In particular, an $n \times n$ matrix is called *vertically symmetric* (VS) if it is “invariant with respect to vertical reflection”, that is, $a_{ij} = a_{i,n+1-j}$ for all i, j . A matrix is called a VSASM (vertically symmetric alternating-sign matrix) if it is VS and ACM at the same time. How many $n \times n$ VSASMs are there? For even n , the VS condition implies that the sum of elements in each row is even, but it should be 1 in any ASM. This contradiction shows that a VSASM may exist only for odd n . Checking the list above, we can see that the matrix $\begin{bmatrix} 1 \end{bmatrix}$ is a VSASM, and the only VSASM of size

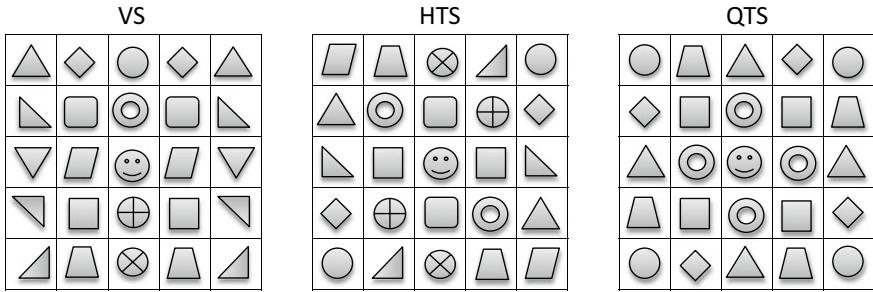


Fig. 2.12 Vertically symmetric (VS), half-turn-symmetric (HTS), and quarter-turn symmetric (QTS) 5×5 matrices

3×3 is $\begin{bmatrix} 0 & 1 & 0 \\ 1 & -1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$. The next possible size is 5×5 , and there are 3 VSASMs of this size, namely,

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & -1 & 1 & 0 \\ 1 & -1 & 1 & -1 & 1 \\ 0 & 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

Counting ASMs with Various Symmetries

There are several other important symmetries a matrix can obey. In particular, an $n \times n$ matrix is called *half-turn symmetric* (HTS) if it is “invariant with respect to 180° rotation”, that is, $a_{ij} = a_{n+1-i, n+1-j}$ for all i, j . Also, an $n \times n$ matrix is called *quarter-turn symmetric* (QTS) if it is “invariant with respect to 90° rotation”, that is, $a_{ij} = a_{j, n+1-i}$, see Fig. 2.12.

The following theorem of Kuperberg [234] counts the number of ASMs in each of these symmetry classes.

Theorem 2.12 *The number of $n \times n$ ASMs is given by*

$$A(n) = (-3)^{\frac{n(n-1)}{2}} \prod_{i=1}^n \prod_{j=1}^n \frac{3(j-i)+1}{j-i+n}.$$

The number of $2n+1 \times 2n+1$ vertically symmetric ASMs (VSASMs) is given by

$$A_V(2n+1) = (-3)^{n^2} \prod_{i=1}^{2n+1} \prod_{j=1}^n \frac{3(2j-i)+1}{2j-i+2n+1}.$$

The number of $2n \times 2n$ half-turn-symmetric ASMs (HTSASMs) is given by

$$A_{HT}(2n) = A(n) \cdot (-3)^{\frac{n(n-1)}{2}} \prod_{i=1}^n \prod_{j=1}^n \frac{3(j-i)+2}{j-i+n}.$$

The number of $4n \times 4n$ quarter-turn symmetric ASMs (QTSASMs) is given by

$$A_{QT}(4n) = A_{HT}(2n) \cdot A(n)^2.$$

For example, let us apply the second formula for $n = 2$, that is, for counting 5×5 VSASMs. Then i takes values from 1 to 5, while j can be 1 or 2, resulting in 10 possible (i, j) pairs. For $i = 1, j = 1$ we have $2j - i = 1$, and the expression $\frac{3(2j-i)+1}{2j-i+2n+1}$ reduces to $\frac{4}{6}$, or $\frac{2}{3}$. We get the same term if $i = 3, j = 2$. Cases $i = 2, j = 1$ and $i = 4, j = 2$ result in the term $\frac{1}{5}$. Continuing this way, we get

$$A_V(5) = (-3)^{2^2} \left(\frac{2}{3}\right)^2 \left(\frac{1}{5}\right)^2 \left(-\frac{1}{2}\right)^2 \cdot \frac{-5}{3} \cdot \frac{7}{7} \cdot \frac{-8}{2} \cdot \frac{10}{8} = 3,$$

which is in line with the direct count.

Further Results

In 1991, David Robbins [323] formulated some conjectures related to counting ASMs obeying various symmetry properties. In addition to vertical symmetry (VS), half-turn-symmetry (HTS), and quarter-turn symmetry (QTC) mentioned above, Robbins asked for the number of ASMs which are (a) vertically and horizontally symmetric, (b) diagonally symmetric, and (c) symmetric in both diagonals. In some cases, Robbins presented explicit conjectural formulas for the number of ASMs in the corresponding symmetry class.

This list of conjectures was like a “roadmap”, stimulating progress in the area. Theorem 2.12 was the first breakthrough in this direction: it resolved the case of vertical symmetry, and also half-turn-symmetry and quarter-turn symmetry for matrices of even size. Later, using the methods developed to prove Theorem 2.12, various researchers were able to prove all the remaining cases! Diagonally symmetric ASMs (with zeros on the diagonal)¹ were counted by Kuperberg himself. Later, Razumov and Stroganov [315, 316] counted half-turn-symmetric and quarter-turn symmetric ASMs for matrices of *odd* size, complementing Theorem 2.12. Vertically and horizontally symmetric ASMs were counted by Okada [292]. Finally, counting ASMs symmetric in both diagonals, the hardest and last case of those listed by Robbins, was done by Behrend, Fischer, and Konvalinka [37].

¹Robbins asked to count the diagonally symmetric ASMs without requiring zeros on the diagonal, but, in this case, he did not get an interesting answer to the conjecture.

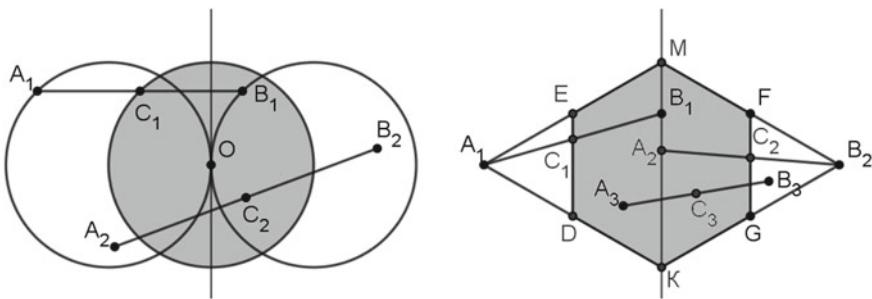


Fig. 2.13 Minkowski symmetrizations of a circle and a triangle

Reference

G. Kuperberg, Symmetry classes of alternating-sign matrices under one roof. *Annals of Mathematics* **156**-3 (2002): 835–866.

2.13 Transforming Convex Bodies into Balls

Minkowski Symmetrization of Circles and Triangles

What does the “average” of you and your mirror image look like? To start with a simple example, let “you” be a circle S with radius 1 and center $(-1, 0)$ in the coordinate plane. Let the “mirror” be the y -axis (that is, the line with equation $x = 0$). The mirror image of S is a circle S' with radius 1 and center $(1, 0)$. Now, we define the “average” of S and S' to be the set $M(S)$ of all possible midpoints C of line segments AB with $A \in S$ and $B \in S'$. In mathematics, $M(S)$ is called the *Minkowski symmetrization* of S with respect to the y -axis. In our case, it is easy to check that $M(S)$ is again a circle with radius 1 and center $(0, 0)$.

Now, let S be the equilateral triangle A_1KM with vertices $A_1(-\sqrt{3}, 0)$, $K(0, -1)$ and $M(0, 1)$. The image of this triangle reflected with respect to the y -axis is the triangle B_2KM , where B_2 has coordinates $(\sqrt{3}, 0)$. In this case one can check that $M(S)$ turns out to be a regular hexagon $DEMFGK$, where D, E, F and G are the midpoints of the sides A_1K , A_1M , B_2M , B_2K , respectively, see Fig. 2.13.

How to Measure “Closeness” to a Circle

Intuitively, a regular hexagon looks “closer” to a circle than a triangle. To make this intuition more precise, let us measure the “closeness” to a circle of any set S in the plane as the ratio $r(S)$ between the radius of the largest circle which can be drawn inside S to the smallest circle containing S . If S is a circle, then this ratio is obviously

1, and this is clearly the largest possible. In general, the closer this ratio is to 1, the “closer” S is to being a circle. For example, if S is a rectangle with sides 1 and 100, the largest circle it contains has radius 0.5, and the smallest circle containing it has radius $\sqrt{(1/2)^2 + (100/2)^2} \approx 50$, hence $r(S) \approx 0.5/50 = 0.01$, in agreement with the intuition that such a rectangle is very far from looking like a circle. If S is an equilateral triangle with side length a , then the radius of the inscribed circle is $\frac{\sqrt{3}}{6}a$ and the radius of the circumscribed circle is $R = \frac{a}{\sqrt{3}}$, hence $r(S) = \frac{1}{2}$, much better. For a square with side length a , we get $r(S) = \frac{a}{2} : \frac{\sqrt{2}a}{2} = \frac{1}{\sqrt{2}} \approx 0.7$. For a regular hexagon with side length a , the radius of the inscribed circle is $\frac{\sqrt{3}}{2}a$ and the radius of the circumscribed circle is a , hence $r(S) = \frac{\sqrt{3}}{2} \approx 0.87$. Because $\frac{\sqrt{3}}{2}$ is much closer to 1 than $\frac{1}{2}$, this formalizes the intuition that Minkowski symmetrization of an equilateral triangle (with respect to the line containing its side) makes it closer to being a circle.

Applying Minkowski Symmetrizations Iteratively

What if we apply Minkowski symmetrization to a regular hexagon, and then again to the resulting image, and so on, will we get a set closer and closer to a circle? How many iterations will we need to make it “sufficiently close”? What if we start not with regular hexagon or triangle, but with some “far from circle-like” set such as a 1×100 rectangle? A theorem of Klartag [227] implies that, no matter where we start from, we can get an “approximate circle” after just 10 iterations!

The same question can be asked in 3-dimensional space. We can start with, say, a regular tetrahedron S , reflect it with respect to any plane to get its image S' , define its Minkowski symmetrization $M(S)$ to be the set of all midpoints of line segments AB with $A \in S$ and $B \in S'$, and ask if $M(S)$ is “closer” to being a ball. The “closeness” of any body S to the ball can be measured as the ratio between the radii of the largest ball inside S (for a regular tetrahedron it is $\frac{a}{\sqrt{24}}$, where a is the side length) and the smallest ball containing S (for a regular tetrahedron it is $\sqrt{\frac{3}{8}}a$, hence the ratio is $1/3$). A theorem of Klartag [227] implies that we can always get “almost a ball” after just 15 iterations of Minkowski symmetrizations!

Convex Bodies and Hyperplanes in n -Dimensional Space

Actually, the theorem below works in *any* dimension, and we need some definitions to formulate it. The n -dimensional Euclidean space \mathbb{R}^n can be described as the set of points with coordinates (x_1, x_2, \dots, x_n) . Given any two points $A = (x_1, x_2, \dots, x_n)$ and $B = (y_1, y_2, \dots, y_n)$, the line segment AB consists of all points with coordinates $(\alpha x_1 + (1 - \alpha)y_1, \alpha x_2 + (1 - \alpha)y_2, \dots, \alpha x_n + (1 - \alpha)y_n)$ for some $\alpha \in [0, 1]$, and

the distance between A and B can be calculated as $|AB| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$. A *ball* with center $O \in \mathbb{R}^n$ and radius $r > 0$ is the set of all points at distance at most r from O . A set $S \subset \mathbb{R}^n$ is called a *convex body*, if it is

- (i) Convex: for any $A, B \in S$ the whole line segment AB belongs to S ,
- (ii) Bounded: there exists a ball B such that $S \subset B$,
- (iii) Closed: for any $A \notin S$ there is a ball with center A which has no intersection with S , and
- (iv) Has non-empty interior: there exists a ball B' such that $B' \subset S$.

The *origin* is the point with coordinates $(0, 0, \dots, 0)$. A *hyperplane* H_a (going through the origin) is the set of points $X = (x_1, x_2, \dots, x_n)$ satisfying the equation

$$\sum_{i=1}^n a_i x_i = 0, \quad (2.25)$$

for some constants a_1, \dots, a_n , not all equal to 0. For example, for $n = 2$ this reduces to $a_1 x + a_2 y = 0$, so in this case a hyperplane is just a line; for $n = 3$ the equation is $a_1 x + a_2 y + a_3 z = 0$, which is the equation of a plane, etc. Equation (2.25) remains the same after multiplication by any non-negative constant, so we can assume without loss of generality that $\sum_{i=1}^n a_i^2 = 1$.

Minkowski Symmetrization in n -Dimensional Space

For any point $X = (x_1, x_2, \dots, x_n)$, its reflection $X' = (x'_1, x'_2, \dots, x'_n)$ with respect to the hyperplane H_a is given by $x'_i = x_i - 2a_i \left(\sum_{j=1}^n a_j x_j \right)$, $i = 1, \dots, n$. For any convex body S , its Minkowski symmetrization $M_a(S)$ is the set of all midpoints of line segments AB with $A \in S$ and B being a reflection of some point in S with respect to the hyperplane H_a . One can check that the Minkowski symmetrization of a convex body is again a convex body.

Theorem 2.13 *Let $n \geq 2$ and let $S \subset \mathbb{R}^n$ be a convex body. Then there exist $5n$ Minkowski symmetrizations M_1, M_2, \dots, M_{5n} such that the convex body*

$$S' = M_{5n}(M_{5n-1}(\dots M_2(M_1(S)) \dots))$$

contains a ball of radius r_1 , but is contained in a ball of radius r_2 , with

$$\frac{r_1}{r_2} = \left(1 - c \frac{|\ln \ln n|}{\sqrt{\ln n}} \right) / \left(1 + c \frac{|\ln \ln n|}{\sqrt{\ln n}} \right) \quad (2.26)$$

where c is a universal constant, independent of n and S .

For $n = 2$, we need only $5n = 10$ iterations to get S' as in Theorem 2.13; for $n = 3$, we need $5n = 15$. For larger n , more iterations are required, but the resulting body S' becomes closer and closer to a ball (the larger n , the smaller the factor $\frac{|\ln \ln n|}{\sqrt{\ln n}}$, hence the ratio $\frac{r_1}{r_2}$ in (2.26) becomes particularly close to 1 in high dimensions).

Interestingly, if S is just a line segment in \mathbb{R}^n , then its Minkowski symmetrization is a line segment again, so no matter how many iterations we do, the result will still be very far from being a ball. Theorem 2.13 is not applicable to the line segment, because it is not a convex body (condition (iv) above fails). For convex bodies, however, it states that surprisingly few iterations suffice.

Reference

- B. Klartag, 5n Minkowski symmetrizations suffice to arrive at an approximate Euclidean ball. *Annals of Mathematics* **156**-3 (2002): 947–960.

Chapter 3

Theorems of 2003



3.1 On the Differentiability of Lipschitz Maps on Infinite-Dimensional Spaces

Functions Which Are “Locally” Almost Linear

What does the graph of the function $y = x^2$ look like around any point, for example, $(1, 1)$? If we look at this point using a stronger and stronger lens, then the graph looks closer and closer to a straight line. Indeed, if $x = 1 + \varepsilon$ for some small $\varepsilon > 0$, then $y = x^2 = 1 + 2\varepsilon + \varepsilon^2 = 1 + 2(x - 1) + \varepsilon^2$. For small ε , ε^2 is negligibly small compared to $1 + 2(x - 1)$, and the graph of the function around $(1, 1)$ looks like the line $y = 1 + 2(x - 1) = -1 + 2x$, plus some small error, see Fig. 3.1. The usual notation for the error is the “little-o notation” o . We write “ $f(x) = o(h(x))$ as $x \rightarrow x_0$ ” if

$$\lim_{x \rightarrow x_0} \frac{f(x)}{h(x)} = 0.$$

With this notation,

$$x^2 = 1 + 2(x - 1) + o(|x - 1|) \quad \text{as } x \rightarrow 1.$$

More generally, if a function $f : \mathbb{R} \rightarrow \mathbb{R}$ “looks like a line” in a neighbourhood of a point $x_0 \in \mathbb{R}$, that is,

$$f(x) = f(x_0) + A(x - x_0) + o(|x - x_0|) \quad \text{as } x \rightarrow x_0, \tag{3.1}$$

for some $A \in \mathbb{R}$, then f is called differentiable at x_0 , and A is called the derivative of f at x_0 .

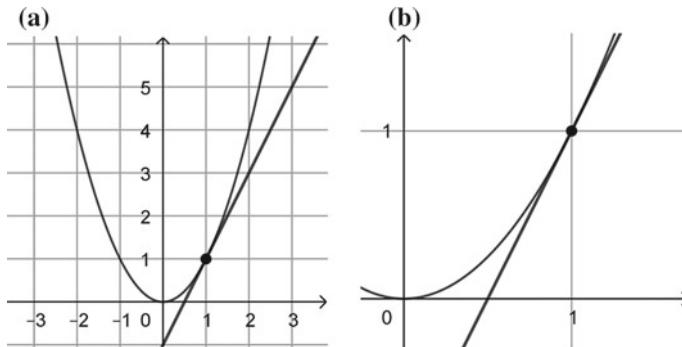


Fig. 3.1 The graph of the function $f(x) = x^2$ is almost a straight line when observed through a magnifier

Locally Linear Functions of Many Variables

Similarly, the graph of the function of two variables $y = f(x_1, x_2) = x_1^2 + 2x_2^2$ looks like a plane in a neighbourhood of any point. For example, $f(1, 1) = 3$, so the point $(1, 1, 3)$ belongs to the graph, and in its neighbourhood

$$y = x_1^2 + 2x_2^2 = 3 + 2(x_1 - 1) + 4(x_2 - 1) + ((x_1 - 1)^2 + 2(x_2 - 1)^2).$$

Because the last term is small, the graph is close to the plane $y = 3 + 2(x_1 - 1) + 4(x_2 - 1)$ in the neighbourhood of $(1, 1, 3)$.

More generally, let $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be a function which maps every point (x_1, x_2) of the plane \mathbb{R}^2 to another point $(y_1, y_2) \in \mathbb{R}^2$, for example f may map (x_1, x_2) to $(x_1^2 + 2x_2^2, x_1 - x_2^2)$. If (x_1, x_2) is close to $(1, 1)$, then $y_1 = x_1^2 + 2x_2^2 \approx 3 + 2(x_1 - 1) + 4(x_2 - 1)$ and similarly $y_2 = x_1 - x_2^2 \approx (x_1 - 1) - 2(x_2 - 1)$, hence the function f is equal to the linear function h mapping the point (x_1, x_2) to the point $(3 + 2(x_1 - 1) + 4(x_2 - 1), (x_1 - 1) - 2(x_2 - 1))$, plus some small error.

A similar result holds even for functions depending on an *infinite* set of variables. Let $x = (x_1, x_2, \dots)$ be any sequence of real numbers converging to 0, and let $f(x) = \max_i x_i^2$. If x is “close” to the point $x_0 = (1, 0, 0, 0, \dots)$, that is, $|x_1 - 1| \leq \varepsilon$, and $|x_i - 0| \leq \varepsilon$, $i = 2, 3, \dots$ for some small $\varepsilon > 0$, then $f(x) = x_1^2 = (1 + (x_1 - 1))^2 \approx 1 + 2(x_1 - 1)$, again approximating f by a linear function.

Some Examples of Banach Spaces

In the examples above, we have met functions $f : X \rightarrow Y$, where X and Y were the set of real numbers, the set of pairs of real numbers, or even sets of sequences converging to 0. All these are examples of *Banach spaces*. Intuitively, a Banach space is a collection of objects, or elements, for which (i) we can define addition and mul-

tiplication by a constant, such that the usual laws hold; (ii) we can define for every $x \in X$ its norm $\|x\|$, again obeying some natural properties like the triangle inequality; and (iii) we can find a limit for every sequence of elements $x_1, x_2, \dots, x_n, \dots$ for which $\|x_{n+1} - x_n\|$ decays sufficiently fast. The set \mathbb{R} of real numbers and the set \mathbb{R}^2 of pairs $x = (x_1, x_2)$ of real numbers are Banach spaces with norms $\|x\| = |x|$ and $\|x\| = \sqrt{x_1^2 + x_2^2}$, respectively. More interestingly, the set c_0 of all sequences $x = (x_1, x_2, \dots, x_n, \dots)$ converging to 0 is also a Banach space: we can add the sequences (or multiply by a number) element-wise, and $\|x\| = \sup_i |x_i|$. The dimension $\dim(X)$ of a Banach space X is, intuitively, the number of coordinates necessary to define any $x \in X$, for example, $\dim(\mathbb{R}) = 1$, $\dim(\mathbb{R}^2) = 2$, and $\dim(c_0) = \infty$. See Sect. 2.11 for the formal definition of a Banach space, more examples, and related discussion.

Locally Linear Functions Between Banach Spaces

To say that a function between Banach spaces is, in a neighbourhood of a point, “almost linear up to a small error”, we need to define what we mean by “linear” and by “small error” in this general setting. For Banach spaces X and Y , a function $T : X \rightarrow Y$ is called *linear* if $T(x + y) = T(x) + T(y)$, $\forall x, y \in X$ and $T(\lambda x) = \lambda T(x)$, for all $\lambda \in \mathbb{R}$, $x \in X$. T is called *bounded* if there exists an $M > 0$ such that $\|T(x)\|_Y \leq M\|x\|_X$, where $\|\cdot\|_X$ and $\|\cdot\|_Y$ are the norms in the spaces X and Y , respectively. A function $g : X \rightarrow Y$ is little-o of the function $h : X \rightarrow \mathbb{R}$ as x approaches x_0 if $\lim_{x \rightarrow x_0} \frac{g(x)}{h(x)} = 0$.

We are finally ready for the following definition. A function $f : X \rightarrow Y$ is called (Fréchet) differentiable at $x_0 \in X$ if there is a bounded linear function $T : X \rightarrow Y$ such that

$$f(x_0 + u) = f(x_0) + T(u) + o(\|u\|) \quad \text{as } u \rightarrow 0.$$

This definition generalizes all the examples above. Needless to say, the ability to approximate a function by a linear one is invaluable in hundreds of applications.

Lipschitz Continuous Functions

Of course, not every function is differentiable, even in one dimension. If $f : \mathbb{R} \rightarrow \mathbb{R}$ is discontinuous at x_0 , its graph around x_0 is not even similar to a line, so differentiability fails. The function $f(x) = |x|$ is continuous everywhere, but is not differentiable at $x_0 = 0$. In 1872, Weierstrass [399] presented an example of a function $f : \mathbb{R} \rightarrow \mathbb{R}$ which is continuous everywhere on \mathbb{R} but differentiable nowhere! Hence, a stronger notion of continuity is necessary to guarantee differentiability at least at some points. A function $f : X \rightarrow Y$ is called *Lipschitz continuous* if there is a constant $K \geq 0$ such that $\|f(x) - f(y)\|_Y \leq K\|x - y\|_X$ for all $x, y \in X$. For example, every

continuous convex¹ function, such as $f(x) = |x|$, is Lipschitz continuous. Every Lipschitz continuous function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is differentiable almost everywhere, that is, the set $A(f)$ of points where it isn't differentiable has Lebesgue measure 0 (can be covered by balls in \mathbb{R}^n with total volume less than ε for any $\varepsilon > 0$). In particular, if f_1, f_2, \dots , is an infinite sequence of such functions, the set $A = A(f_1) \cup A(f_2) \cup \dots$ has Lebesgue measure 0, hence $A \neq \mathbb{R}^n$, and there is a point x such that all f_i are simultaneously differentiable at x .

Simultaneous Differentiability of Lipschitz Continuous Functions

However, there is no notion of “volume” or “Lebesgue measure” in infinite-dimensional space, and therefore the argument above cannot be applied to analyse the differentiation of functions $f : X \rightarrow Y$ for infinite-dimensional Banach spaces X and Y . In particular, there were no methods to prove the existence of a point of simultaneous differentiability even for two functions f_1, f_2 defined on any Banach space X with $\dim(X) = \infty$, until the following theorem was proved in 2003 by Lindenstrauss and Preiss [245].

Theorem 3.1 *There exist Banach spaces X and Y with $\dim(X) = \infty$, and a family of subsets of X called the Γ -null sets, such that (i) every Lipschitz continuous function $f : X \rightarrow Y$ is Fréchet differentiable outside a Γ -null set, and (ii) any countable union of Γ -null sets is again a Γ -null set, and, in particular, cannot cover the whole X . In particular, (i) and (ii) imply that for any sequence f_1, f_2, \dots , of Lipschitz continuous functions $X \rightarrow Y$ there exists a point $x \in X$ at which they are all simultaneously differentiable.*

In fact, Lindenstrauss and Preiss established explicit sufficient conditions for X and Y which guarantee that Theorem 3.1 holds, and used them to prove the theorem for many specific examples of X and Y . In particular, it works for $X = c_0$ and $Y = \mathbb{R}$, as well as for some examples when Y is also infinite-dimensional.

Reference

J. Lindenstrauss and D. Preiss, On Fréchet differentiability of Lipschitz maps between Banach spaces. *Annals of Mathematics* **157**-1 (2003): 257–288.

¹A function $f : X \rightarrow Y$ is called *convex* if $f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$ for all $x \in X, y \in Y$, and for all $\alpha \in [0, 1]$.

3.2 Representing 1 as a Sum of Reciprocals of Selected Integers

Dividing Pizzas Using Egyptian Fractions

How would you divide 5 pizzas among 6 diners? Each one should get $\frac{5}{6}$ of a pizza, hence we may divide each pizza into 6 equal pieces, resulting in 30 pieces in total, and then give 5 pieces to each diner. This method, however, requires a lot of cutting work. Alternatively, we may notice that $\frac{5}{6} = \frac{1}{2} + \frac{1}{3}$, divide 3 pizzas into halves, the remaining 2 pizzas into thirds, and then give each diner one half and one third. With this method, the total number of pieces reduces from 30 to just 12 (Fig. 3.2)!

This is only one out of numerous applications where it is convenient to represent a rational number as a sum of unit fractions: $\frac{a}{b} = \sum_{i=1}^n \frac{1}{x_i}$ where x_i are positive integers.

To simplify the notation, we may write $[x_1, x_2, \dots, x_n]$ instead of $\sum_{i=1}^n \frac{1}{x_i}$, hence, for example, $\frac{5}{6}$ can be written as just $[2, 3]$. Such representations were studied in ancient Egypt, and are also called *Egyptian fractions*. In fact, Egyptian fractions were used as the standard notation for rational numbers for many centuries, until the modern notation ($\frac{a}{b}$, or decimal) was developed.

Representations of 1 and the Erdős–Graham Conjecture

Of special interest are representations of 1 in the form $\sum_{i=1}^n \frac{1}{x_i}$, with all positive integers x_i being distinct. The simplest one is $1 = \frac{1}{2} + \frac{1}{3} + \frac{1}{6}$, or $1 = [2, 3, 6]$ in short. Replacing $\frac{1}{3}$ by $\frac{1}{4} + \frac{1}{12}$, we get another representation $1 = [2, 4, 6, 12]$, where only even numbers are used in the denominators. Can we find a representation where all denominators are odd and greater than 1? This requires a bit more work to find, but also exists, the simplest one being $1 = [3, 5, 7, 9, 11, 15, 35, 45, 231]$.

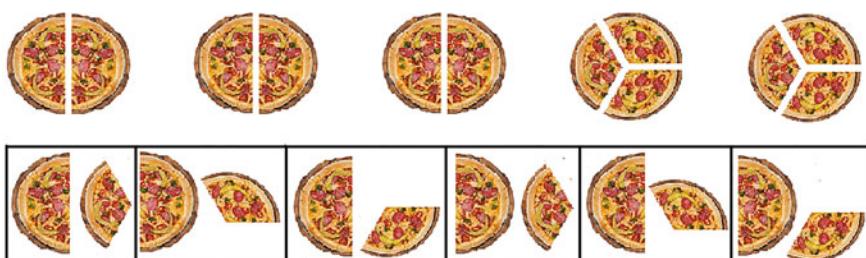


Fig. 3.2 Division of 5 pizzas among 6 diners with just 12 pieces

Hence, if we divide all integers into two classes, even and odd, then denominators taken from only one class (in this case, from either one) are sufficient to represent 1 as a sum of distinct unit fractions. Is this always the case? For example, let us divide the integers into another two classes: composite numbers and primes. Can we use only primes as denominators? It turns out, we cannot. If

$$1 = \sum_{i=1}^n \frac{1}{p_i},$$

where p_i are distinct primes, then $M = \sum_{i=1}^n \frac{M}{p_i}$, or $M - \sum_{i=2}^n \frac{M}{p_i} = \frac{M}{p_1}$, where $M = p_1 p_2 \dots p_n$. However, M is divisible by p_1 and each $\frac{M}{p_i}$ for $i \neq 1$ is divisible by p_1 as well, hence $M - \sum_{i=2}^n \frac{M}{p_i}$ is divisible by p_1 . On the other hand, $\frac{M}{p_1}$ is not divisible by p_1 , a contradiction. Hence, we cannot use only primes to represent 1. However, we can use only composite numbers, for example, $1 = [4, 6, 8, 10, 12, 15, 18, 20, 24, 30, 36]$.

In 1980, Erdős and Graham [135] asked if this is the case in general, that is, for any division of the integers $2, 3, 4, \dots$ into $r \geq 2$ classes, can we always represent 1 as a sum of distinct unit fractions using denominators from one class only?

An Attempt to Build a Counterexample

To get an idea of how the answer can be negative, let us solve a similar question about the representation of $\frac{61}{144}$ in the form $\frac{61}{144} = \sum_{i=1}^n \frac{1}{x_i^2}$, where $x_i \geq 2$ are distinct integers. In general, this is possible: $\frac{61}{144} = \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2}$. However, let us put 2 and 3 in the first class, and all other integers in the second one. Then $\frac{1}{2^2} + \frac{1}{3^2} < \frac{61}{144}$, hence the equality is impossible with x_i from the first class. If x_1, \dots, x_n are in the second class,

$$\sum_{i=1}^n \frac{1}{x_i^2} \leq \sum_{i=4}^{n+3} \frac{1}{i^2} \leq \sum_{i=4}^{n+3} \frac{1}{i(i-1)} = \sum_{i=4}^{n+3} \left(\frac{1}{i-1} - \frac{1}{i} \right)$$

$$= \frac{1}{3} - \frac{1}{4} + \frac{1}{4} - \frac{1}{5} + \dots + \frac{1}{n+2} - \frac{1}{n+3} = \frac{1}{3} - \frac{1}{n+3} < \frac{1}{3} < \frac{61}{144},$$

hence the equality is impossible with x_i from the second class as well. In each class, the sum is always less than the number we are trying to reach!

For similar reasons, we could answer the Erdős–Graham question negatively if we could divide the integers into two classes such that the sum of the reciprocals of the integers in each class is less than 1. Let us try to do this. Because $\frac{1}{2} + \frac{1}{3} < 1$, we can put 2 and 3 into the first class. Then, $\frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \frac{1}{9} < 1$, so we can safely put 4, 5, 6, 7, 8, 9 into the second class. Now, $\frac{1}{2} + \frac{1}{3} + \frac{1}{10} < 1$, so 10 can go

to the first class, but wherever we put 11, the corresponding sum becomes greater than 1. In fact, $\sum_{i=2}^{11} \frac{1}{i} > 2$, so numbers from 2 to 11 cannot be divided into two classes with sum of reciprocals less than 1 in each class.

Counterexample: The Search for More Than Two Classes

What if we can use $r \geq 3$ classes? Then 11 can go to the third one, but, for any r , there is a number $N(r)$ such that

$$\sum_{i=2}^{N(r)} \frac{1}{i} > r, \quad (3.2)$$

hence numbers from 2 to $N(r)$ cannot be divided into r classes with sum of reciprocals less than 1 in each class. For example, (3.2) holds with $N(r) = 4^r$, because

$$\sum_{i=2}^{4^r} \frac{1}{i} = \sum_{k=0}^{r-1} \left(\sum_{i=2^{2k}+1}^{2^{2(k+1)}} \frac{1}{i} \right) > \sum_{k=0}^{r-1} \left(\sum_{i=2^{2k}+1}^{2^{2(k+1)}} \frac{1}{2^{2(k+1)}} \right) = \sum_{k=0}^{r-1} \frac{1}{2} = r.$$

Using a similar argument, one can show that $\sum_{i=2}^{2^r} \frac{1}{i} < r$, so we can hope to divide integers from 2 to 2^r into r classes with sum of reciprocals being less than 1 in each class. However, we cannot divide *all* integers into r classes in this way.

The Proof of the Erdős–Graham Conjecture

Obviously, one can try other methods for dividing integers into classes. For example, let $r = 2$, let us put 2 in the first class (call it A), and 3 in the second class (B), to be sure that the elements of the triple $[2, 3, 6]$ are not in the same class. Because $1 = [2, 4, 6, 12]$, let us put 4 in class B as well. Because $1 = [3, 4, 5, 6, 20]$, these numbers should not all be in the same class B , so let us put 5 in A . Also, $1 = [3, 4, 6, 10, 12, 15]$, so let us put 6 in A as well. Now, $1 = [2, 6, 7, 12, 14, 28]$, and 2 and 6 are in A , so let us put 7 in B . Continuing in this way, we may choose 10, 14, 15 to be in B , 20 and 21 in A , but then we would have a problem with 28: putting it into A would result in a representation $1 = [2, 5, 6, 20, 21, 28]$ (all denominators are in A), while putting it into B would result in $1 = [3, 4, 7, 10, 14, 15, 28]$ (all in B). However, this particular issue can be “corrected” by other choices in the earlier stages (for example, why not put both 2 and 3 in A , and 4, 5, 6 in B and so on), and more classes can be used, so where is the warranty that there is no smart way to divide *all* integers into r classes such that we *never* get a contradiction like with 28 above? The following theorem of Croot [106] provides such a warranty.

Theorem 3.2 *There exists a constant b such that for every partition of the integers in $[2, b^r]$ into r classes, there is always one class containing a subset S with the property $\sum_{i \in S} \frac{1}{i} = 1$.*

Theorem 3.2 not only gives a positive answer to the Erdős–Graham question, but provides an explicit bound where any potential counterexample is guaranteed to stop working. In fact, Croot shows that, for sufficiently large r , $b = 10^{72527}$ suffices. As explained above, Theorem 3.2 does not hold if we replace the interval $[2, b^r]$ in it by the interval $[2, 2^r]$. Hence, the exponential dependence on r is unavoidable, and the bound b^r in Theorem 3.2 is optimal up to the value of b .

Reference

E.S. Croot III, On a coloring conjecture about unit fractions. *Annals of Mathematics* **157**-2 (2003): 545–556.

3.3 The Optimal Hardy–Littlewood Maximal Inequality

The Integral as an Area Below the Graph of a Function

How would you calculate the area below a curve on the coordinate plane? For example, what is the area of the region

$$S = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq x^2\}. \quad (3.3)$$

More generally, for any non-negative function f on $[a, b]$, what is the area below the plot of f , that is, of the region $\{(x, y) : a \leq x \leq b, 0 \leq y \leq f(x)\}$? The standard notation for this area is

$$\int_a^b f(x)dx, \quad (3.4)$$

which is called an *integral*. For example, if f is constant on $[a, b]$, that is, $f(x) = C$ for every $x \in [a, b]$, then $\int_a^b f(x)dx = \int_a^b Cdx = C(b - a)$, using the formula for the area of a rectangle. The area of the set S in (3.3) can then be written as

$$\int_0^1 x^2 dx,$$

but how do we calculate this integral, or, more generally, the integral (3.4)?

How to Calculate an Area/Integral?

To calculate (3.4), we introduce the function $F(t) = \int_a^t f(x)dx$, which has properties

(i) $F(a) = \int_a^a f(x)dx = 0$ and (ii) for any $\varepsilon > 0$,

$$F(t + \varepsilon) - F(t) = \int_t^{t+\varepsilon} f(x)dx. \quad (3.5)$$

The last integral denotes the area of the set $S' = \{(x, y) : t \leq x \leq t + \varepsilon, 0 \leq y \leq f(x)\}$. Intuitively, if ε becomes smaller and smaller, this set becomes “closer and closer” to the rectangle with sides $(t + \varepsilon) - t = \varepsilon$ and $f(t)$, whose area is $\varepsilon f(t)$. Hence, $\frac{1}{\varepsilon} \int_t^{t+\varepsilon} f(x)dx$ is approximately equal to $f(t)$. This approximation does not hold exactly, but the smaller ε is, the better it holds. In such cases mathematicians write

$$\lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \int_t^{t+\varepsilon} f(x)dx = f(t). \quad (3.6)$$

Substituting (3.6) into (3.5) results in

$$\lim_{\varepsilon \rightarrow 0} \frac{F(t + \varepsilon) - F(t)}{\varepsilon} = f(t). \quad (3.7)$$

The left-hand side of (3.7) is called the *derivative* of $F(t)$. Hence, to calculate the integral (3.4), it is sufficient to find a function F with $F(a) = 0$ whose derivative is equal to f . In particular, the derivative of $F(t) = t^3/3$ is ${}^2 t^2$, and $F(0) = 0^3/3 = 0$, hence $\int_0^t x^2 dx = \frac{t^3}{3}$, and the area of the set S in (3.3) is $\int_0^1 x^2 dx = \frac{1^3}{3} = \frac{1}{3}$.

Finite Areas of Some Infinite Regions

The same method can be used to calculate areas of *infinite* regions, and the resulting area is often a finite number. For example, the area of the infinite region $\{(x, y) : 1 \leq x, 0 \leq y \leq 1/x^2\}$ is the integral

$$\int_1^\infty \frac{1}{x^2} dx. \quad (3.8)$$

The derivative of the function $F(t) = 1 - \frac{1}{t}$ is $\lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \left(1 - \frac{1}{t+\varepsilon} - 1 + \frac{1}{t}\right) = \lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \frac{\varepsilon}{t(t+\varepsilon)} = \frac{1}{t^2}$, and $F(1) = 1 - \frac{1}{1} = 0$, hence the integral in (3.8) is equal to $F(t)$ evaluated at $t = \infty$. It is clear that $\frac{1}{t}$ becomes 0 for $t = \infty$, hence the answer is 1. The set of all functions $f : \mathbb{R} \rightarrow \mathbb{R}$ such that the integral

²If you are not familiar with the formulas for derivatives, you can verify this directly from definition (3.7): $\lim_{\varepsilon \rightarrow 0} \frac{(t+\varepsilon)^3/3 - t^3/3}{\varepsilon} = \frac{1}{3} \lim_{\varepsilon \rightarrow 0} \frac{t^3 + 3t^2\varepsilon + 3t\varepsilon^2 + \varepsilon^3 - t^3}{\varepsilon} = \frac{1}{3} \lim_{\varepsilon \rightarrow 0} (3t^2 + 3t\varepsilon + \varepsilon^2) = t^2$.

$$\int_{-\infty}^{+\infty} |f(x)|dx$$

exists and finite, is denoted $\mathcal{L}^1(\mathbb{R})$, and the value of this integral is denoted $\|f\|_1$.

The Lebesgue Differentiation Theorem

Integrals are used not only for calculating areas, but everywhere in mathematics, and the described method is central for calculating them. However, our intuitive justification of (3.6) is not a formal proof. In fact, (3.6) may even be wrong: consider a function $f(t)$ which is 0 for $t \neq 0$ but $f(0) = 1$. Then, for any ε , $\int_0^\varepsilon f(x)dx = 0$, hence $\lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \int_0^\varepsilon f(x)dx = 0 \neq 1 = f(0)$, that is, (3.6) is wrong for $t = 0$. However, for this function, this is the only exception: for any $t \neq 0$ (3.6) holds. The classical Lebesgue differentiation theorem states that this is true in general: for any f (for which the integral can be defined) a version of (3.6), namely,

$$\lim_{\varepsilon \rightarrow 0} \frac{1}{2\varepsilon} \int_{t-\varepsilon}^{t+\varepsilon} f(x)dx = f(t), \quad (3.9)$$

holds for “almost all” points t (formally, it states that the set of points for which (3.9) may fail has “measure 0”, that is, it can be covered by a collection of intervals of total length δ for any $\delta > 0$).

The Hardy–Littlewood Maximal Inequality

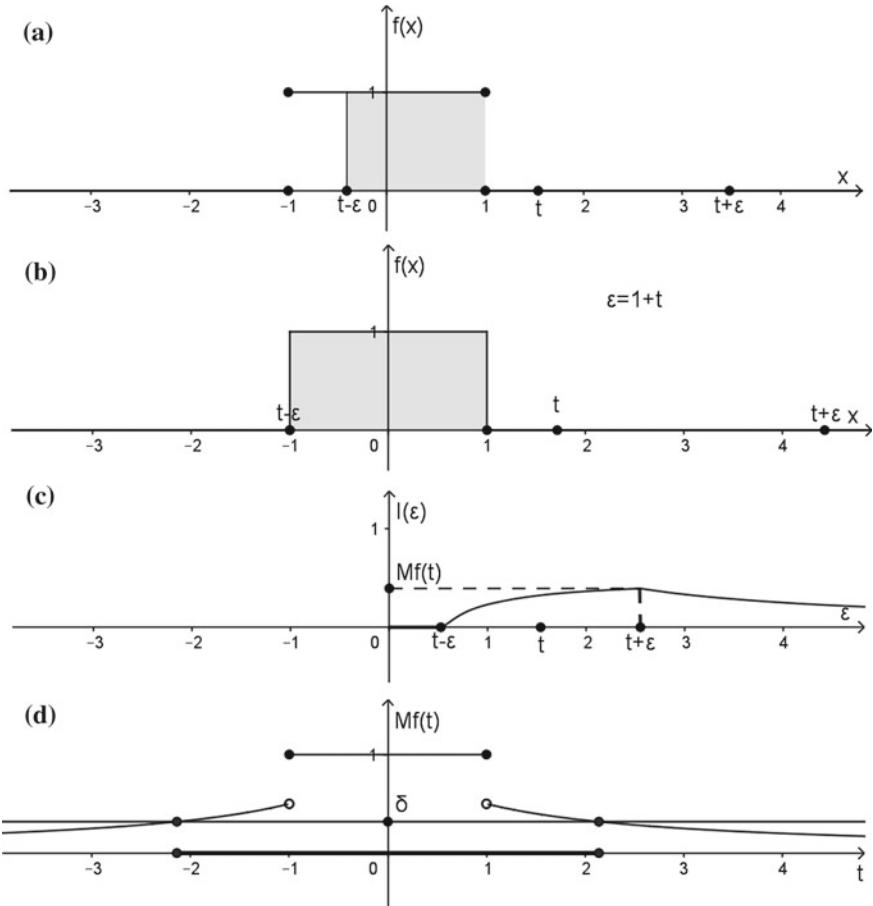
To prove this theorem, one needs to understand how big the integral in (3.9) can be. For any function $f \in \mathcal{L}^1(\mathbb{R})$, define

$$M_f(t) = \sup_{\varepsilon > 0} \frac{1}{2\varepsilon} \int_{t-\varepsilon}^{t+\varepsilon} |f(x)|dx,$$

where the right-hand side means the smallest real number which is greater than or equal to $\frac{1}{2\varepsilon} \int_{t-\varepsilon}^{t+\varepsilon} |f(x)|dx$ for all $\varepsilon > 0$.

For example, let $f(x) = D$ (for some $D > 0$) for $-1 \leq x \leq 1$, and $f(x) = 0$ otherwise. Then for any $t \geq 1$ the expression $I(\varepsilon) := \frac{1}{2\varepsilon} \int_{t-\varepsilon}^{t+\varepsilon} |f(x)|dx$ is maximal for $\varepsilon = 1 + t$, see Fig. 3.3a–c, and is equal to $\frac{D}{1+t}$. In other words, $M_f(t) = \frac{D}{1+t}$ for $t \geq 1$. Similarly, $M_f(t) = \frac{D}{1-t}$ for $t \leq -1$, and $M_f(t) = D$ for $t \in [-1, 1]$, see Fig. 3.3d.

The function $M_f(t)$ converges to 0 as t goes to $+\infty$ or $-\infty$. In other words, for any $\delta > 0$, we have $M_f(t) < \delta$ for all t except on the finite interval $t \in [1 - \frac{D}{\delta}, \frac{D}{\delta} - 1]$ of length $2(\frac{D}{\delta} - 1) < \frac{2D}{\delta}$, see Fig. 3.3d. Denoting by $|\{M_f > \delta\}|$ the length of this exceptional interval, and noting that $2D = \int_{-\infty}^{+\infty} |f(x)|dx = \|f\|_1$, we

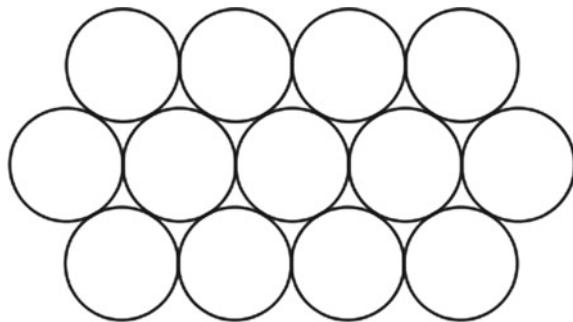
Fig. 3.3 Construction of $M_f(t)$

get an estimate $|\{M_f > \delta\}| < \frac{\|f\|_1}{\delta}$. The famous centered Hardy–Littlewood maximal inequality [148] states that this estimate is *always* true, up to a constant. That is, there is a constant C such that

$$|\{M_f > \delta\}| \leq C \frac{\|f\|_1}{\delta} \quad (3.10)$$

holds true for all functions $f \in \mathcal{L}^1(\mathbb{R})$ and all $\delta > 0$. Using this, one can formally prove the Lebesgue differentiation theorem (3.9) and go ahead with calculating integrals.

Fig. 3.4 Hexagonal circle packing



What is the Best Constant C in (3.10)?

The Hardy–Littlewood maximal inequality has also a lot of other applications, and in some of them it is important to know what is the best possible constant C such that (3.10) holds. This question was open for decades, until was finally resolved by Melas [269].

Theorem 3.3 *For every $f \in \mathcal{L}^1(\mathbb{R})$ and for every $\delta > 0$ we have*

$$|\{M_f > \delta\}| \leq \frac{11 + \sqrt{61}}{12} \frac{\|f\|_1}{\delta}, \quad (3.11)$$

and the constant $\frac{11 + \sqrt{61}}{12} \approx 1.5675208$ in (3.11) is the best possible.

Reference

A.D. Melas, The best constant for the centered Hardy–Littlewood maximal inequality. *Annals of Mathematics* **157**-2 (2003): 647–688.

3.4 Improved Upper Bounds on Sphere Packings

People with Umbrellas and Circle Packing in the Plane

How many people can attend a concert which takes place in large park with area S ? If each person occupies area s , we have an upper bound of S/s . However, allowing this many people would be problematic, they would fill the entire park without any holes, could not walk, could not open an umbrella in case of rain, etc. Let us be more polite and assume that each person occupies, say, a circle of radius r , where r is sufficient to open an umbrella. Such a circle occupies area $s = \pi r^2$, but now the upper bound S/s cannot be achieved, because circles cannot fit together to occupy the whole area without holes.

So, how can we arrange equal circles in the plane as densely as possible? The answer is easy to guess after a bit of experimentation. Each circle can have at most 6 neighbours which touch it, whose centres form a regular hexagon with side length $2r$. Continuing in this way, we obtain the so-called *hexagonal packing arrangement* (Fig. 3.4), which occupies a portion of about $\frac{\pi}{2\sqrt{3}} \approx 0.9069$ of the entire area. In 1890, Axel Thue proved that this is optimal, see [86] for a simplified proof of this result.

Error Correcting Codes and Sphere Packing in Higher Dimensions

The same problem naturally arises in higher dimensions. For example, what is the densest way to pack oranges into a large box, how many oranges we can fit in? If oranges are modelled as spheres of equal size, this is exactly the question of finding the densest packing of equal spheres in 3-dimensional space. In dimensions higher than 3, we have no oranges, but the problem is still important, and has a connection, for example, to error correcting codes. Imagine you need to transfer a long message via a noisy channel, and you would like your recipient to be able to recover the *exact* message you sent. For this, you can encode your message in a digital form, and divide it into parts of equal length n , so that each part $x_1x_2\dots x_n$ can be considered as a point $x = (x_1, x_2, \dots, x_n)$ in n -dimensional space. Now, assume that recipient receives it in perturbed form as $y_1y_2\dots y_n$, such that the distance $\|x - y\| = \sqrt{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2}$ between the points corresponding to the original and the perturbed messages is less than r . How can he or she recover the $x_1x_2\dots x_n$ from $y_1y_2\dots y_n$? In general, this is impossible, but imagine that you have promised in advance that your message has some special property, say, all x_i are either 0 or 1 and $\sum_{i=1}^n x_i 2^{i-1}$ is a power of 3. Let S be the set of all messages with this property. Then, after receiving message $y_1y_2\dots y_n$, the recipient can find $x \in S$ which is the closest to y , and guess that this was the original message. If the distance between any two elements of S is not less than $2r$, then, for any y , there will in fact be at most one $x \in S$ at distance less than r , so this guess is guaranteed to be correct. The only disadvantage with this approach is that we can now send only $|S|$ different messages of length n , where $|S|$ is the number of elements in S . With our choice of S , $|S| < n$, because there are only a few powers of 3 less than 2^n , but can we do better? All we need is a set S of points in n -dimensional space³ such that (i) the distance between any 2 elements of S is not less than $2r$, and (ii) $|S|$ is as large as possible. Because (i) can be reformulated as “spheres with centres in S and radii r do not intersect”, this is exactly the problem of finding a densest possible sphere packing in n -dimensional space!

³In fact, if we insist that all x_i are either 0 or 1, then we could select only $S \subset S'$, where S' is the set of points with this property, but this problem is still similar to sphere packing.

The Center Density of Sphere Packing

Let us formulate this problem more rigorously. We can assume that $r = 1$, that is, we will pack unit spheres. A set S of points in \mathbb{R}^n such that $\|x - y\| \geq 2$ for all distinct $x, y \in S$ is called a *sphere packing*, and elements of S correspond to the centres of the spheres. The *center density* of a sphere packing S is

$$\delta(S) := \limsup_{R \rightarrow \infty} \frac{|S|}{|B_n(0, R)|},$$

where $B_n(0, R)$ is the ball with centre $0 = (0, 0, \dots, 0) \in \mathbb{R}^n$ and radius R , and $|B_n(0, R)|$ is the n -dimensional volume of this ball. The problem is then to find a sphere packing S with largest possible $\delta(S)$. Let δ_n be the optimal center density in dimension n .

Lower and Upper Bounds for Optimal Packing Density

It is relatively easy to find a lower bound for δ_n : for this, we just need to provide a good example of a dense packing. For example, in $n = 8$, let S' be the set of points $x = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)$ such that either all x_i are integers, or all x_i are halves of odd integers, and $\sum_{i=1}^8 x_i$ is an even integer. Then, for any distinct $x, y \in S'$ either $|x_i - y_i| \geq 0.5$, $i = 1, \dots, 8$ or there are at least two indices $i, j \in 1, \dots, 8$ such that $|x_i - y_i| \geq 1$ and $|x_j - y_j| \geq 1$, in either case $\|x - y\| \geq \sqrt{2}$. Then $S = \{x \mid x = \sqrt{2}y, y \in S'\}$ is a sphere packing. One can check that S' has on average one point per unit cube, hence S has on average one point per cube with side length $\sqrt{2}$ and volume $(\sqrt{2})^8 = 16$, thus $\delta(S) = \frac{1}{16}$, which gives the lower bound $\delta_8 \geq \frac{1}{16}$.

Proving the upper bounds is much trickier. In order to prove an upper bound $\delta_n \leq B$, we need to prove that $\delta(S) \leq B$ for *any* n -dimensional sphere packing S . Before 2003, the best upper bound for δ_8 was within the factor 1.01216 from the lower bound, and the situation was worse in all other dimensions $n \geq 4$.

Upper Bounds Using the Fourier Transform

In 2003, Henry Cohn and Noam Elkies [94] developed a beautiful technique which allows us to significantly improve the upper bounds in all dimensions $n \geq 4$, and, more importantly, the method can be used for further improvements, or even for the complete solution of the problem in certain dimensions. To formulate their bound, we need the notion of the *Fourier transform*. For any integrable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, its Fourier transform $\hat{f} : \mathbb{R}^n \rightarrow \mathbb{C}$ is defined as

$$\hat{f}(t) := \int_{\mathbb{R}^n} f(x) e^{2\pi i \langle x, t \rangle} dx,$$

where \mathbb{C} is the set of complex numbers (see e.g. Sect. 1.7), $i = \sqrt{-1}$, $t = (t_1, t_2, \dots, t_n)$ is a vector, $\langle x, t \rangle := \sum_{i=1}^n x_i t_i$ is the scalar product on \mathbb{R}^n , and $dx = dx_1 dx_2 \dots dx_n$, so that $\hat{f}(t)$ is the n -dimensional integral of a complex function. A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is called *admissible* if there exist positive constants ε , C_1 , C_2 such that $|f(x)| \leq \frac{C_1}{(1+\|x\|)^{n+\varepsilon}}$ and $|\hat{f}(x)| \leq \frac{C_2}{(1+\|x\|)^{n+\varepsilon}}$ for all $x \in \mathbb{R}^n$.

Theorem 3.4 Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is an admissible function, is not identically zero, and satisfies the following two conditions:

- (i) $f(x) \leq 0$ for $\|x\| \geq 1$, and
- (ii) $\hat{f}(t) \geq 0$ (which implies that $\hat{f}(t)$ is real) for all $t \in \mathbb{R}^n$.

Then the center density δ_n of n -dimensional sphere packings is bounded above by $\frac{f(0)}{2^n \hat{f}(0)}$.

In all dimensions $n \geq 4$, Cohn and Elkies provided examples of functions f such that all conditions of Theorem 3.4 hold, and the corresponding upper bounds are better than previously known ones. In particular, for $n = 8$, this gives an upper bound within the ratio just 1.000001 from the lower bound, while in dimension $n = 24$ the corresponding ratio is 1.0007071, so they come amazingly close to the complete resolution of the sphere packing problem in these dimensions. More importantly, Theorem 3.4 opens the road for further improvements: one “just” needs to find an example of a function f satisfying all the conditions of the theorem with ratio $\frac{f(0)}{2^n \hat{f}(0)}$ as low as possible.

In fact, Maryna Viazovska [392] later found a function f in dimension $n = 8$ which, when substituted into Theorem 3.4, gives an upper bound *exactly* equal to the lower bound, thus fully solving the sphere packing problem in dimension 8. After this, a group of mathematicians [96] used Theorem 3.4 to fully solve the problem in dimension $n = 24$ as well!

Reference

H. Cohn and N. Elkies, New upper bounds on sphere packings I. *Annals of Mathematics* **157**-2 (2003): 689–714.

3.5 Sums Versus Products of Finite Sets of Integers

The Set $S(A)$ of All Possible Sums

Let A be the set of distinct integers with $|A| = k$ elements, and

$$S(A) = \left\{ \sum_{i=1}^k \varepsilon_i a_i \mid a_i \in A, \varepsilon_i = 0 \text{ or } 1 \right\}$$

be the set of all possible sums of the elements of A . How many distinct integers can $S(A)$ contain? Because each ε_i is either 0 or 1, there are 2^k sums $\sum_{i=1}^k \varepsilon_i a_i$, and, if they are all different, then $S(A)$ contains $|S(A)| = 2^k$ elements, and this is clearly the largest possible. Can all these sums really be different? It turns out that they can. For example, this is the case if $A = \{1, 2, 4, 8, \dots, 2^{k-1}\}$ is the set of powers of 2. Indeed, in this case, consider $s = \sum_{i=1}^k \varepsilon_i 2^{i-1}$ and $s' = \sum_{i=1}^k \varepsilon'_i 2^{i-1}$, and let i be the largest index such that $\varepsilon_i \neq \varepsilon'_i$. Then, either $\varepsilon_i = 1$ and $\varepsilon'_i = 0$, or vice versa. In the first case, $s \geq 2^{i-1} > 2^{i-1} - 1 = \sum_{j=1}^{i-1} 2^{j-1} \geq s'$, and similarly $s' > s$ in the second case. Hence, in either case, $s \neq s'$. For example, the set $\{1, 2, 4\}$ has $2^3 = 8$ subsets – the empty set $\{\}$, the sets $\{1\}$, $\{2\}$, $\{4\}$, $\{1, 2\}$, $\{1, 4\}$, $\{2, 4\}$, and $\{1, 2, 4\}$, and you can directly check that the sums of elements in all these subsets are indeed different.

On the other hand, for $A = \{1, 2, \dots, k\}$, we have $0 \leq s \leq \sum_{i=1}^k i = \frac{1}{2}k(k+1)$ for any $s \in S(A)$, hence $|S(A)|$ is at most $\frac{1}{2}k(k+1) + 1$, which for large k is a tiny number compared to 2^k . More generally, if A is any *arithmetic progression*, that is, a set of the form

$$A = \{a, a+b, a+2b, \dots, a+(k-1)b\},$$

for some integers a and $b \neq 0$, then any $s \in S(A)$ can be written in the form $s = xa + yb$, where $0 \leq x \leq k$ and $0 \leq y \leq \sum_{i=1}^{k-1} i = \frac{1}{2}k(k-1)$, hence we have $|S(A)| \leq (k+1) \cdot \frac{1}{2}k(k-1) < k^3$, again a tiny number compared to 2^k . For example, for $k = 100$, k^3 is just a million, while 2^k is a 31-digit number.

The Set $\Pi(A)$ of All Possible Products

Similar to $S(A)$, we can study the set

$$\Pi(A) = \left\{ \prod_{i=1}^k a_i^{\varepsilon_i} \mid a_i \in A, \varepsilon_i = 0 \text{ or } 1 \right\}$$

of all possible *products* of elements of A . If all a_i are different prime numbers, then the uniqueness of prime factorization implies that all 2^k products are distinct and $|\Pi(A)| = 2^k$. For $A = \{1, 2, \dots, k\}$, there are about $\frac{k}{\ln k}$ primes in A , and therefore $|\Pi(A)| \geq 2^{k/\ln k}$, again a very large number. In particular, as a function of k , it grows faster than any polynomial, even with huge degree like k^{1000} . More precisely, for any degree d , there exists a constant $k_0 = k_0(d)$ such that $2^{k/\ln k} > k^d$ for any $k \geq k_0(d)$. Indeed, $2^{k/\ln k} > k^d$ is equivalent to $k/\ln k > d \log_2 k$, or $k > d \ln k \log_2 k$, which is obviously true for large enough k .

However, for $A = \{1, 2, 4, 8, \dots, 2^{k-1}\}$, any element $m \in \Pi(A)$ is a power of 2, that is, $m = 2^x$, where $0 \leq x \leq \sum_{i=1}^{k-1} i = \frac{1}{2}k(k-1)$, hence $|\Pi(A)| \leq \frac{1}{2}k(k-1) + 1 < k^2$. More generally, if A is any *geometric progression*, that is, a set of the form

$$A = \{a, aq, aq^2, \dots, aq^{k-1}\},$$

for some integers $a \neq 0$ and q with $|q| \geq 2$, then any $m \in \Pi(A)$ can be written in the form $m = a^x q^y$, where $0 \leq x \leq k$ and $0 \leq y \leq \sum_{i=1}^{k-1} i = \frac{1}{2}k(k-1)$, hence $|\Pi(A)| < k^3$.

The Sum-Product Conjecture

The attentive reader has probably already observed an interesting phenomenon: if $|S(A)|$ is small, such as in the case $A = \{1, 2, \dots, k\}$, then $|\Pi(A)|$ is large, while when $|\Pi(A)|$ is small, such as in the case $A = \{1, 2, 4, 8, \dots, 2^{k-1}\}$, then $|S(A)|$ is large. Can we construct for any k a k -element set A such that *both* $|S(A)|$ and $|\Pi(A)|$ are small, say, of a size bounded by some polynomial in k ? In 1983, Erdős and Szemerédi [139] conjectured that this is *not* the case, and, for any set A , either the set of sums $S(A)$ or the set of products $\Pi(A)$ (or both) is very large:

Conjecture 3.1 *Let*

$$g(k) := \min_{A: |A|=k} (|S(A)| + |\Pi(A)|). \quad (3.12)$$

Then $g(k)$ grows faster than any polynomial in k , that is, for any d , there exists a constant $k_0 = k_0(d)$ such that $g(k) > k^d$ for any $k \geq k_0(d)$.

The idea behind Conjecture 3.1 is the following. To have a small $|S(A)|$, the set A should have a special structure with respect to addition, like an arithmetic progression, which would result in almost all out of 2^k possible sums being the same. Similarly, to have a small $|\Pi(A)|$, the set A should have some special structure with respect to multiplication, like a geometric progression. Conjecture 3.1 then states that addition and multiplication are such different operations that there exist no sets which are “special” with respect to both addition and multiplication at the same time.

Upper and Lower Bounds for $g(k)$

Erdős and Szemerédi [139] proved an *upper* bound for $g(k)$, namely, they proved that there exists a constant c such that

$$g(k) < e^{c \frac{(\ln k)^2}{\ln(\ln k)}},$$

where $e \approx 2.72$ is the base of the natural logarithm, and all logarithms in the formula are natural. However, to prove an upper bound $g(k) < h(k)$ it is sufficient, for every k , to provide an example of an A such that $|S(A)| + |\Pi(A)| < h(k)$. In contrast, Conjecture 3.1 asks for a *lower* bound for $g(k)$, but proving a lower bound in the form $g(k) > f(k)$ requires a proof that *no* set A can have $|S(A)| + |\Pi(A)| \leq f(k)$. Because there are so many methods for selecting A , proving that none of them works

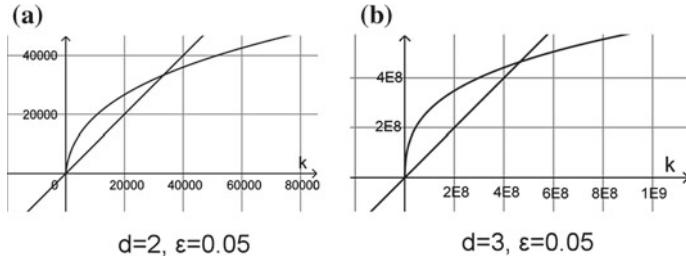


Fig. 3.5 The inequality $k > (\ln k)^{\frac{d}{1/2-\varepsilon}}$ holds for large enough k

is much more difficult. There was essentially no progress for 20 years, until Mei-Chu Chang [87] gave a complete answer.

Theorem 3.5 *Let $g(k)$ be given by (3.12). Then there is an $\varepsilon > 0$ such that*

$$k^{\left(\frac{1}{2}-\varepsilon\right)\frac{\ln k}{\ln(\ln k)}} < g(k) < k^{\left(1+\varepsilon\right)\frac{\ln k}{\ln(\ln k)}}.$$

The main contribution in Theorem 3.5 is of course the lower bound. To see that Theorem 3.5 implies Conjecture 3.1, we need to verify the inequality

$$k^{\left(\frac{1}{2}-\varepsilon\right)\frac{\ln k}{\ln(\ln k)}} > k^d.$$

This is equivalent to $(\frac{1}{2}-\varepsilon)\frac{\ln k}{\ln(\ln k)} > d$, or $\ln k > \frac{d}{1/2-\varepsilon} \ln(\ln k) = \ln(\ln k)^{\frac{d}{1/2-\varepsilon}}$, which in turn is equivalent to $k > (\ln k)^{\frac{d}{1/2-\varepsilon}}$. Because $\frac{d}{1/2-\varepsilon}$ is just a fixed constant, this inequality is true for large enough k , see Fig. 3.5.

Because the lower and upper bounds in Theorem 3.5 are very similar, the theorem not only proves that $g(k)$ grows faster than any polynomial (Conjecture 3.1), but establishes (almost) exactly *how* fast it grows.

Reference

M.-C. Chang, The Erdős–Szemerédi problem on sum set and product set. *Annals of Mathematics* **157**–3 (2003): 939–957.

3.6 The Radius of Integer Points in the Plane

Customer Phone Calls and Poisson Processes

Assume that you work in a company, your job is to answer customers' phone calls, and customers call you at random times, on average D customers per hour, with all calls being independent from each other. The (random) number $X(\lambda)$ of customers who call you after λ hours is called a *Poisson process*, and this is the most popular model for counting customers' calls, car accidents (they happen at random times, independently

from each other, on average D accidents per week), claims that insurance companies receive, and thousands of similar situations.

Counting Pairs of Calls Within a Minute From Each Other

Assume that you need at least a minute to answer any call, so if two customers called you within a minute from each other, then you are in trouble and need to ask your colleagues to help you. How often will this happen? Assume that the first customer called you at time λ_1 , measured in hours. Then you are in trouble if someone else calls you within the next minute, that is, between times λ_1 and $\lambda_1 + \frac{1}{60}$. During the next hour, you expect about D customers, so you would guess that only about $D/60$ of them will call within the next minute, and will “intersect” with the first customer. By a similar argument, $D/60$ of customers could intersect with the second customer, and so on. Because during the next λ hours you expect about λD customers, you would expect in total about $(\lambda D)D/60$ pairs of them to call within one minute from each other.

Of course, this is just an estimate, which is never exactly true, because customers call you at random times. Sometimes you may be lucky to have significantly fewer such “collisions”, and sometimes significantly more. However, it is hard to expect that you can be continuously lucky, say, during a year. So, if $r_{1/60}(\lambda)$ is the number of pairs of customers who call you within a minute after λ hours, you would expect that our estimate $r_{1/60}(\lambda) \approx (\lambda D)D/60$ will work reasonably well at least if λ is large. More formally,

$$\lim_{\lambda \rightarrow \infty} \frac{r_{1/60}(\lambda)}{\lambda D} = \frac{D}{60},$$

with probability 1.

Counting Pairs of Calls at “Distance” Between a and b

Of course, there is nothing special about *one* minute here. By a similar argument, after λ hours, you would expect about $(\lambda D)Db$ customers to call within b hours from each other, and about $(\lambda D)Da$ customers within a hours. Assuming $a < b$ and subtracting, you can conclude that there will be about $(\lambda D)D(b-a)$ pairs of customers such that the “time distance” between their calls is greater than a but less than b . More formally, if $\lambda_1 \leq \lambda_2 \leq \dots$ are the exact times of customer calls, $|\{j \neq k \mid \lambda_j \leq \lambda, \lambda_k \leq \lambda, a \leq \lambda_j - \lambda_k \leq b\}|$ is the number of pairs of customers who call before time λ and the distance between their calls belong to $[a, b]$, and

$$R_2[a, b](\lambda) := \frac{1}{D\lambda} |\{j \neq k \mid \lambda_j \leq \lambda, \lambda_k \leq \lambda, a \leq \lambda_j - \lambda_k \leq b\}|, \quad (3.13)$$

then

$$\lim_{\lambda \rightarrow \infty} R_2[a, b](\lambda) = D(b - a) \quad (3.14)$$

with probability 1.

Distances From a Fixed Point to All Points with Integer Coordinates in the Plane

In 1991, motivated by some deep physical theory, Cheng and Lebowitz [89] performed the following experiment. They chose any point X with coordinates (α, β) on the coordinate plane, and measured (squares of the) distances from X to all points $Y_{m,n}$ whose coordinates (m, n) are integers. Then they ordered the resulting numbers

$$(m - \alpha)^2 + (n - \beta)^2$$

in non-decreasing order to form an infinite sequence $\lambda_1 \leq \lambda_2 \leq \dots$, see Fig. 3.6, and performed some numerical experiments to understand the properties of this sequence. While this sequence is completely deterministic (that is, does not involve any randomness), they found that, amazingly, the properties of this sequence are very similar to the properties of the (random) sequence of λ 's denoting the exact times of

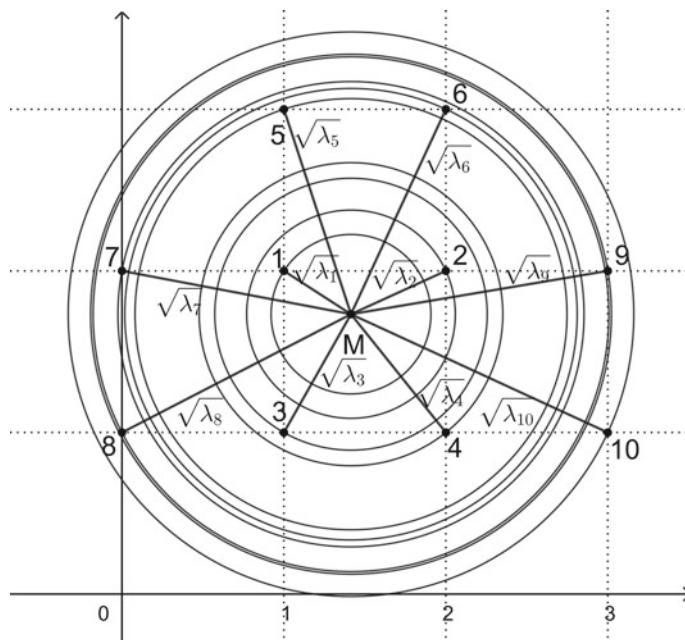


Fig. 3.6 Distances from fixed point M to all points with integer coordinates

customer calls, if the average rate of calls is $D = \pi$. They observed this phenomenon for almost any initial point X they chose, with a few exceptions.

Explaining the Coincidence

Well, the average rate $D = \pi$ can be easily explained. For large λ , the number of j such that $\lambda_j \leq \lambda$ is equal to the number of points with integer coefficients (m, n) such that $(m - \alpha)^2 + (n - \beta)^2 \leq \lambda$, or, in other words, to the number of integer points in a disk with center X and radius $r = \sqrt{\lambda}$. This number can be approximated by the area of the disk, see Sect. 1.3, which is $\pi r^2 = \pi\lambda$. This is exactly the expected number of customers who will call before time λ if the average rate of calls is π customers per hour.

However, this easy geometric argument does not explain why this sequence of λ 's, arising from distances in the coordinate plane, also satisfies the deeper properties of Poisson processes, like (3.14). Nevertheless, the following theorem of Marklof [261] proves that this is indeed the case.

Theorem 3.6 *Suppose that $\alpha, \beta, 1$ are linearly independent over \mathbb{Q} , and assume α is diophantine. Let $\lambda_1 \leq \lambda_2 \leq \dots$ be defined as above (that is, equal to the numbers of the form $(m - \alpha)^2 + (n - \beta)^2$ for integer m, n , arranged in non-decreasing order), let $a < b$ be real numbers, and $R_2[a, b](\lambda)$ be given by (3.13). Then*

$$\lim_{\lambda \rightarrow \infty} R_2[a, b](\lambda) = \pi(b - a).$$

Diophantine Numbers

In Theorem 3.6, \mathbb{Q} is the set of rational numbers, and the condition “ $\alpha, \beta, 1$ are linearly independent over \mathbb{Q} ” implies that there are no rational numbers x, y, z , (except $x = y = z = 0$), such that $x\alpha + y\beta + z = 0$. This eliminates the case when α or β are rational, as well as cases like $\alpha = \sqrt{2}, \beta = 2\sqrt{2} + 3$.

An irrational number α is called *diophantine* if for every $\varepsilon > 0$ there exists a $C_\varepsilon > 0$ such that

$$\left| \alpha - \frac{p}{q} \right| \geq \frac{C_\varepsilon}{q^{2+\varepsilon}}$$

for every pair of integers p and $q \neq 0$. Intuitively, α is diophantine if it is not a rational number, and, moreover, it is “reasonably far” from any rational number $\frac{p}{q}$. One may show that Theorem 3.6 does not hold if α and β are rational, or irrational and well approximable by rationals, hence the condition “ α is diophantine” in the theorem cannot be omitted. This condition excludes some artificial examples of irrational numbers like, e.g., Liouville's constant, $L = \sum_{j=1}^{\infty} 10^{-j!}$, but it is well known that “almost all” real numbers are diophantine. In particular, all irrational algebraic numbers (that is, roots of non-zero polynomials with rational coefficients) are diophantine.

For example, the pair $\alpha = \sqrt{2}$ and $\beta = \sqrt{3}$ satisfy all the conditions of Theorem 3.6: indeed, they are roots of polynomials $x^2 - 2$ and $x^2 - 3$, respectively, hence $\sqrt{2}$ and $\sqrt{3}$ are algebraic, and therefore diophantine. Also, $x\sqrt{2} + y\sqrt{3} + z = 0$ for rational x, y, z is possible only if $x = y = z = 0$. Hence, Theorem 3.6 tells us that, if we measure the distances from all points with integer coefficients to $X = (\sqrt{2}, \sqrt{3})$, square them, and arrange them in non-decreasing order, the resulting sequence would behave like the times of customer calls: for large λ , the sequence will contain about $\pi\lambda$ terms less than λ , and, for any $a < b$, about $(\lambda\pi)\pi(b - a)$ pairs of indices j, k such that $\lambda_j \leq \lambda, \lambda_k \leq \lambda$, and $a \leq \lambda_j - \lambda_k \leq b$. If you are seeing this theorem for the first time, it may seem like an almost impossible, “magical” connection between seemingly unrelated areas of mathematics. For specialists who see the “big picture”, this result is probably much less surprising, and even expected.

Reference

J. Marklof, Pair correlation densities of inhomogeneous quadratic forms. *Annals of Mathematics* **158**-2 (2003): 419–471.

3.7 The Set of Nonergodic Directions Can Have Dimension 1/2

The Movement of a Billiard Ball

One of the central objects of study in mathematics are *dynamical systems*, that is, systems evolving in time according to some fixed rule. For example, the system of n bodies moving under gravitation is a dynamical system, which is, however, extremely difficult to describe for $n \geq 3$, so mathematicians need simpler models to start with. One of the simplest examples of a dynamical system is the motion of a billiard ball. Assume that there is only one ball on a rectangular table, there is no friction, so that the ball, after an initial hit, will move at a constant speed, with the usual rule that the angle of incidence equals the angle of reflection. The ball is modelled as a point, and it may either reach a vertex and stop there, or move forever. What can we say about the “typical” trajectory, that is, starting from a random point and random direction? Of course, for any initial point, there are some directions such that the ball reaches a vertex and stops, and some directions leading to simple-to-describe periodic trajectories (for example, parallel to the sides of a rectangle). However, a classical result of Kerckhoff, Masur, and Smillie [218] (KMS for short) from 1986 implies that for “almost every” initial direction, the resulting trajectory will be “equidistributed” throughout the whole table.

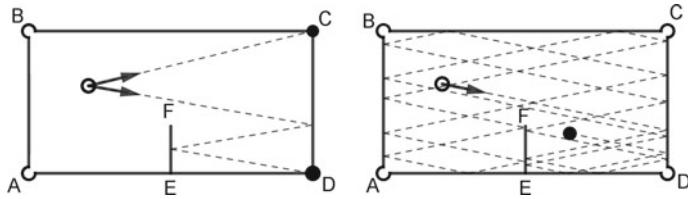


Fig. 3.7 Non-ergodic directions (left) and an ergodic direction (right) in a rational polygon

Equidistributed Trajectories and Ergodic Directions

What does “equidistributed” mean? Intuitively, it means that the ball trajectory “covers” the whole table “uniformly”. For example, let us divide the whole rectangular table into two equal sub-rectangles, M_1 and M_2 . If the trajectory is “equidistributed”, the ball will spend about “half” of the time in M_1 , and “half” of the time in M_2 . For example, the ball may start at time $t = 0$ from a point inside M_1 . Then, at some time $t_1 > 0$ the ball moves to M_2 , spends some time in it, moves back to M_1 at time $t_2 > t_1$, then again to M_2 at some time $t_3 > t_2$, and so on. Then the total time the ball spent inside M_1 during the time interval $[0, T]$ is $f_1(T) := t_1 + (t_3 - t_2) + (t_5 - t_4) + \dots$, while the time spent inside M_2 is $f_2(T) := (t_2 - t_1) + (t_4 - t_3) + \dots$, where the summation continues until we reach time T . By definition, $f_1(T) + f_2(T) = T$, and, for an “equidistributed” trajectory, $f_1(T) \approx f_2(T) \approx T/2$ for large T . More formally, $\lim_{T \rightarrow \infty} \frac{f_1(T)}{T} = \lim_{T \rightarrow \infty} \frac{f_2(T)}{T} = \frac{1}{2}$. Similarly, if the table is divided into three equal parts, an “equidistributed” ball will spend about third of the time inside each of them. More generally, for any part M of the table Q for which the area $S(M)$ is well-defined, the ball is expected to spend about a portion $\frac{S(M)}{S(Q)}$ of its time inside M . Formally, the trajectory is called *equidistributed* if

$$\lim_{T \rightarrow \infty} \frac{f_M(T)}{T} = \frac{S(M)}{S(Q)}, \quad (3.15)$$

where $f_M(T)$ is the time the ball spent inside M during the time interval $[0, T]$.

In fact, the KMS Theorem holds not only for the usual rectangular table, but for any table in the form of a rational polygon, that is, a polygon whose angles measured in radians are rational multiples of π . Rational polygons may be non-convex, have angles greater than π , and even repeated vertices and sides. For example, Fig. 3.7 depicts a rational 7-gon $BCDEFEA$ with sides AB , BC , CD , DE , EF , FE , and EA , and angles $\angle EAB = \angle ABC = \angle BCD = \angle CDE = \angle FEA = \angle ABC = \pi/2$, but $\angle EFE = 2\pi$.

The directions leading to trajectories satisfying (3.15) are also called *ergodic*, while all other directions are called *nonergodic*, and the set of all nonergodic directions in a polygon Q is denoted $NE(Q)$. In this terminology, the KMS Theorem states that almost every initial direction is ergodic.

The Lebesgue Measure of the Set of Nonergodic Directions

The KMS Theorem states that (3.15) holds for “almost every” initial direction. However, what exactly is meant by “almost every”? After all, the set $NE(Q)$ of nonergodic directions is infinite: we can select a vertex A and find directions such that (0) the ball goes directly to A , (1) the ball goes to A after 1 reflection, (2) after 2 reflections, and so on. All these trajectories are finite and cannot satisfy (3.15), hence all these directions belong to $NE(Q)$. However, mathematicians have developed a lot of tools to show that even infinite sets may be in some sense “small”. The most popular “indicator of smallness” is when a set has Lebesgue measure 0. A subset C of an interval $[a, b]$ has Lebesgue measure 0 if, for any $\varepsilon > 0$, it can be covered by intervals of total length less than ε . In Sect. 1.4 we show that, for example, the set of all rational numbers in $[0, 1]$ has measure 0, and, moreover, we study an example of a set, called the Cantor set, which has “as many points as the whole interval $[0, 1]$ ”, but still has Lebesgue measure 0. Now we can finally formulate the KMS result [218] completely formally: it states that, for every rational polygon Q , the set $NE(Q)$ of nonergodic directions, when considered as a subset of the interval $[0, 2\pi]$ of all possible directions, has Lebesgue measure 0.

The Hausdorff Dimension of a Set

This, however, is not the end of the story. In 1992, Masur [263] proved an even stronger upper bound: in fact, the set $NE(Q)$ has not only measure 0, but dimension at most $1/2$! What does this mean, and how can the dimension be fractional?

In fact, one of the intuitive interpretations of dimension is how the “volume” of an object scales with its size. To cover the unit interval, we need 2 intervals of length $1/2$; to cover the unit square, we need $2^2 = 4$ squares with side length $1/2$; to cover the unit cube, we need $2^3 = 8$ cubes with side length $1/2$. In general, we need about $N_\varepsilon = (1/\varepsilon)^d$ cubes with side lengths ε to cover the d -dimensional unit cube. From this, we can find the dimension d as $d = \ln(N_\varepsilon)/\ln(1/\varepsilon)$.

In general, we can cover “larger” sets by “smaller” sets of arbitrary form, not necessary by cubes. For any set B , its diameter $D(B)$ is the smallest real number r such that the distance $\|x - y\|$ is at most r for any $x, y \in B$. Now, for any set S , let N_ε be the minimal number of sets of diameter at most ε sufficient to cover S . Then, informally,⁴ the Hausdorff dimension of S is

$$d := \lim_{\varepsilon \rightarrow 0} \frac{\ln(N_\varepsilon)}{\ln(1/\varepsilon)}. \quad (3.16)$$

⁴In general, the limit (3.16) may not exist, so the formal definition is more complicated. If $A = (A_1, A_2, \dots)$ is a collection of closed sets that cover the set S , let $C^d(A, S) = \sum_i D(A_i)^d$, and let $C^d(S) = \inf_A C^d(A, S) = \inf_A \sum_i D(A_i)^d$, where the infimum is taken over all possible collections A of closed sets that cover set S . Then the *Hausdorff dimension* of S is $\dim(S) := \inf\{d \geq 0 : C^d(S) = 0\}$.

In other words, the Hausdorff dimension is the number d such that $N_\varepsilon \approx (1/\varepsilon)^d$ for small ε .

The Dimension Can Be Fractional!

In (3.16), we do not require d to be an integer. For example, imagine that some set $S \subset [a, b]$ is so small that it can be covered by just 2 intervals of length $1/4$ each, or by 10 intervals of length $1/100$ each, or, more generally, by just $N_\varepsilon = \sqrt{1/\varepsilon}$ intervals of length ε each, for any $\varepsilon > 0$. In this case, N_ε grows as $(1/\varepsilon)^d$ with $d = 1/2$, and we say that such a set S is $1/2$ -dimensional.

In fact, for any real number $r \geq 0$ there exist sets with Hausdorff dimension r . If $\dim(S) < 1$, then it is not difficult to show that S has Lebesgue measure 0. But different sets of measure 0 can have different Hausdorff dimension: for example, the dimension of the set of rational numbers is 0, but the dimension of the (classical) Cantor set described in Sect. 1.4 is $\frac{\ln 2}{\ln 3} \approx 0.63$. By modifying the parameters of that construction, we may build a Cantor-like set of dimension $1 - \varepsilon$ for any $\varepsilon > 0$, and by making a union of such sets we may even construct a set of measure 0 but Hausdorff dimension 1.

Hence, Masur's 1992 result that $NE(Q)$ has Hausdorff dimension at most $1/2$ is a much stronger statement than the 1986 KMS Theorem stating that it has measure 0. However, maybe $NE(Q)$ is even smaller than this. Could even stronger upper bounds on its Hausdorff dimension be possible?

The Exact Value of the Hausdorff Dimension of $NE(Q)$

The following theorem of Cheung [90] fully answers this question by providing an example of a polygon Q with $\dim(NE(Q))$ exactly equal to $1/2$. For any diophantine⁵ number $\lambda \in (0, 1)$, let Q_λ be the polygon $BCDEFEA$ depicted in Fig. 3.7, with side lengths $|AB| = 1$, $|BC| = 2$, and $|EF| = \lambda$.

Theorem 3.7 *For any diophantine number $\lambda \in (0, 1)$, the Hausdorff dimension $\dim(NE(Q_\lambda))$ is equal to $\frac{1}{2}$.*

Theorem 3.7 shows that Masur's 1992 upper bound of $1/2$ is the best possible, and therefore fully resolves the question of how big the dimension of the set $NE(Q)$ of nonergodic directions in a rational polygon Q can be.

⁵A number λ is called *diophantine* if for every $\varepsilon > 0$ there exists a $C_\varepsilon > 0$ such that $\left| \lambda - \frac{p}{q} \right| \geq \frac{C_\varepsilon}{q^{2+\varepsilon}}$ for every rational number $\frac{p}{q}$, see Sect. 3.6 for a more detailed discussion.

Further Research and Open Questions

While Theorem 3.7 is a big step forward, we still do not know the answers to many natural questions about the movement of a billiard ball in the described model. For example, we do not know if there exists a rational billiard table such that its set of nonergodic directions has Hausdorff dimension strictly between 0 and $\frac{1}{2}$. Also, in addition to understanding *how many* nonergodic directions there are, it is interesting to understand how exactly the ball will move if pushed in such a direction. The simplest example of “nonergodic” movement is a periodic trajectory, that is, one which repeats itself over and over again. There is a 200-year-old conjecture stating that every triangular billiard table contains at least one such trajectory. Thanks to a breakthrough result of R. Schwartz [340], we now know that this conjecture is true for every obtuse triangle⁶ with angle at most 100 degrees. However, the general case remains wide open.

Reference

Y. Cheung, Hausdorff dimension of the set of nonergodic directions. *Annals of Mathematics* **158**-2 (2003): 661–678.

3.8 A Real Number Which Is Far from All Cubic Algebraic Integers

The Impossibility of an Exact Measurement of the Diagonal of the Unit Square

It was noticed in ancient times that the diagonal of a unit square cannot be measured exactly, whatever precise unit of measurements you use. If the length of the square side is 1 m, then, in meters, the diagonal length is between 1 and 2, in decimetres between 14 and 15, in centimetres between 141 and 142, in millimetres between 1414 and 1415, and so on, never being an exact integer amount in any unit of measurements, see Fig. 3.8.

Indeed, from Pythagoras’ theorem we know that the diagonal length is $\sqrt{2}$ metres. Assume that this is equal to exactly a/b -metres, where “ b -meter” is the unit of measurement dividing 1 m into b equal parts. In this case $\sqrt{2}$ could be written as a fraction $\frac{a}{b}$ for some integers a and b . If a and b have some common divisors, they can be cancelled out, and we can write $\frac{a}{b}$ as an irreducible fraction $\frac{a'}{b'}$, for example $\frac{8}{12} = \frac{2 \cdot 4}{2 \cdot 6} = \frac{4}{6} = \frac{2 \cdot 2}{2 \cdot 3} = \frac{2}{3}$. Then $\sqrt{2} = \frac{a}{b} = \frac{a'}{b'}$, which implies that $(b')^2 = 2(a')^2$. Hence, $(b')^2$ is an even number, which implies that b' is an even number as well, and can be written as $b' = 2c$ for some integer c . Then $(2c)^2 = 2(a')^2$, hence $(a')^2 = 2c^2$ is an even number, which implies that a' is an even number as well. But if a' and b' are both even, the fraction $\frac{a'}{b'}$ cannot be irreducible, which is a contradiction.

⁶A triangle is called obtuse if it has an obtuse angle, that is, an angle larger than 90 degrees.

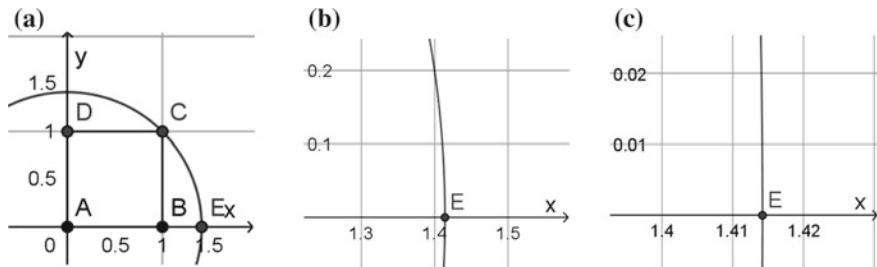


Fig. 3.8 The length of $AE = AC$ cannot be measured exactly

Approximate Measurements of the Diagonal

While the exact representation of $\sqrt{2}$ as $\frac{a}{b}$ is impossible, can we at least *approximate* $\sqrt{2}$ in such a way? Of course we can, with any precision we want. For example, to find a, b such that $\left| \sqrt{2} - \frac{a}{b} \right| < 0.005$, we can choose $b = 100$, and select the integer a minimizing $\left| \sqrt{2} - \frac{a}{b} \right|$, which is $a = 141$, and results in $\left| \sqrt{2} - \frac{141}{100} \right| \approx 0.0042$. Similarly, we can guarantee the inequality $\left| \sqrt{2} - \frac{a}{b} \right| < \varepsilon$ for any $\varepsilon > 0$ by selecting $b > \frac{1}{2\varepsilon}$. However, it would be nicer to have a good approximation as “simple” as possible, and “non-simplicity” of approximation can be measured as (the absolute value of) b . For example, the fraction $\frac{17}{12}$ approximates $\sqrt{2}$ almost twice as well as $\frac{141}{100}$ ($\left| \sqrt{2} - \frac{17}{12} \right| < 0.0025$), and this approximation is also “nicer” because the denominator is much smaller.

Rational Approximation of Irrational Numbers

Numbers presentable in the form $\frac{a}{b}$ for integers a and $b \neq 0$ are called rational, while all other real numbers are called irrational. A theorem of Borel [68] from 1903 states that for every irrational number α there are infinitely many rational numbers $\frac{a}{b}$ such that

$$\left| \alpha - \frac{a}{b} \right| < \frac{1}{\sqrt{5}b^2}. \quad (3.17)$$

In other words, if we would like the approximation error $\left| \alpha - \frac{a}{b} \right|$ to be less than some specific $\varepsilon > 0$, we should expect the denominator b to be such that $\frac{1}{\sqrt{5}b^2} \approx \varepsilon$, which results in $b \approx \sqrt{(\sqrt{5}\varepsilon)^{-1}}$. For $\varepsilon = 0.0025$ this gives $b \approx 13$, which implies that finding the approximation $\frac{17}{12}$ to $\sqrt{2}$ was not just good luck. However, for an approximation with $\varepsilon = 10^{-10}$ we need $b \approx \sqrt{(\sqrt{5}\varepsilon)^{-1}} \approx 67000$, and fractions with such denominators are not as “nice” and “simple” as $\frac{17}{12}$. Can Borel’s theorem be

improved? Can we find good approximations with even smaller denominators? Unfortunately, this is not the case, and there are numbers, for example $\phi = \frac{1+\sqrt{5}}{2}$, which is called *golden ratio*, for which the estimate (3.17) in Borel's theorem is tight and cannot be improved.

Approximation by Expressions Involving Square Roots

However, why approximate ϕ by *rational* numbers if we have a nice *exact* formula for it using $\sqrt{5}$? Obviously, for some irrational numbers, like $\pi \approx 3.1415\dots$, no similar formula exists, but why not use formulas with square roots for approximation? For example, $\pi \approx \sqrt{10}$ approximates π better than $\pi \approx 3$. However, while we evaluate the simplicity of approximation $\frac{a}{b}$ as the absolute value of b (the smaller $|b|$ the better), how do we compare the “simplicity” of expressions, say, $\frac{1+\sqrt{5}}{2}$ and $\sqrt{10}$? For this, we note that $\phi = \frac{1+\sqrt{5}}{2}$ is a root of the nice equation $\phi^2 - \phi - 1 = 0$, while the simplest equation with root $x = \sqrt{10}$ is $x^2 - 10 = 0$. In general, if there is a polynomial P with integer coefficients such that $P(\alpha) = 0$, let us call the *height* $H(\alpha)$ of the number α the largest absolute value of the coefficients of P . Obviously, ϕ is also a solution of the equation $100\phi^2 - 100\phi - 100 = 0$, so we need to select a polynomial which gives $H(\alpha)$ as small as possible. For example, $H(\phi) = 1$, while $H(\sqrt{10}) = 10$.

We say that $\alpha \in \mathbb{R}$ is an *algebraic integer* of degree d if $P(\alpha) = 0$ for some polynomial P of degree d with integer coefficients and leading coefficient 1. So, how well can we approximate any irrational number ξ by algebraic integers of degree 2, such as ϕ or $\sqrt{10}$? In 1969, Davenport and Schmidt [109] proved that there is a constant C such that for any irrational ξ there are infinitely many algebraic integers α of degree at most 2 such that

$$0 < |\xi - \alpha| < \frac{C}{H(\alpha)^2},$$

and that this is the best possible up to the constant factor. You can see that the dependence of the approximation error on $H(\alpha)$ is inverse-quadratic, the same as the dependence on b in (3.17). So, it does not seem that we have made significant progress by allowing numbers like ϕ or $\sqrt{10}$ to be used in the approximation.

Approximation by Cubic Algebraic Integers

However, why stop at algebraic integers of degree at most 2, and not use numbers like $\sqrt[3]{10}$ in the approximation? In this case, we can finally do better! Let ξ be any real number that cannot be written exactly in the form $\frac{a+b\sqrt{c}}{d}$ for some integers a, b, c, d . Then Davenport and Schmidt proved that there are infinitely many algebraic integers α of degree at most 3 such that

$$|\xi - \alpha| < \frac{C}{H(\alpha)^{\gamma^2}}, \quad (3.18)$$

for some constant $C \in \mathbb{R}$. Here, $\gamma^2 = \left(\frac{1+\sqrt{5}}{2}\right)^2 = \frac{3+\sqrt{5}}{2} \approx 2.618 > 2$, so we have finally got better than inverse-quadratic dependence. Now, we can expect approximation of any ξ within an error of order $\varepsilon \approx 10^{-10}$ by algebraic integers with height about $10^{10/\gamma^2} \approx 10^{3.82} \approx 6600$, a significant improvement compared to $10^{10/2} = 10^5$. However, the exponent γ^2 here looks a bit strange, is it optimal or can we do better? A natural conjecture was that we can, and in fact we can achieve approximation in the form $|\xi - \alpha| < \frac{C}{H(\alpha)^3}$ by algebraic integers α of degree at most 3, and, more generally, approximation in the form

$$|\xi - \alpha| < \frac{C}{H(\alpha)^n},$$

by algebraic integers α of degree at most n .

Surprisingly, this natural conjecture turned out to be false, even for $n = 3$, as the following theorem of Damien Roy [328] states.

Theorem 3.8 *There exists a transcendental number ξ and a constant $c_1 > 0$ such that, for any algebraic integer α of degree at most 3, we have*

$$|\xi - \alpha| \geq \frac{C_1}{H(\alpha)^{\gamma^2}}.$$

Here, “transcendental” means that ξ is not a root of any polynomial (of any degree) with rational coefficients. Hence, it is a bit hard to explicitly write down this magic number ξ which is so difficult to approximate. However, we now know that such a number exists. Hence, the exponent in (3.18) cannot be improved from γ^2 to 3, and in fact estimate (3.18) is optimal up to the constant factor.

Reference

D. Roy, Approximation to real numbers by cubic algebraic integers II. *Annals of Mathematics* **158**-3 (2003): 1081–1087.

Chapter 4

Theorems of 2004



4.1 The Julia Set of Almost All Quadratic Polynomials is Locally Connected

When is the Sequence $x, f(x), f(f(x)), \dots$ Bounded?

For a function $f : \mathbb{R} \rightarrow \mathbb{R}$ and initial point $x_0 \in \mathbb{R}$, consider the sequence of points

$$x_0, \quad x_1 = f(x_0), \quad x_2 = f(x_1) = f(f(x_0)), \quad \dots, \quad x_n = f(x_{n-1}), \quad \dots$$

It turns out that, even if f is as simple as a piecewise-linear function or quadratic polynomial, this sequence can behave in an extremely complicated way, and even the simplest questions about it, like “For what values of x_0 is the sequence bounded?” can be highly non-trivial and lead to very interesting mathematics.

For example, in Sect. 1.4 we showed that for the function

$$f(x) = \begin{cases} 3x, & \text{if } x \leq 0.5, \\ 3(1-x), & \text{if } x \geq 0.5, \end{cases}$$

the set K_f of x_0 for which the sequence is bounded is a Cantor set, that is, the set remaining after we remove from the interval $[0, 1]$ the middle third $(1/3, 2/3)$, then remove the middle third from both remaining intervals $[0, 1/3]$ and $[2/3, 1]$, then from the 4 remaining intervals, and so on up to infinity. After step n , we are left with 2^n intervals of length $1/3^n$ each, so that the “length” of K_f is $\lim_{n \rightarrow \infty} (2/3)^n = 0$, but, as we have seen in Sect. 1.4, the set K_f is non-empty, and in fact contains as many points as the whole interval $[0, 1]$.

Functions of a Complex Variable and Julia Sets

Problems of this type are usually studied for functions $f : \mathbb{C} \rightarrow \mathbb{C}$, where \mathbb{C} is the set of *complex* numbers, that is, numbers of the form $a + ib$, where $a, b \in \mathbb{R}$, and $i = \sqrt{-1}$. We can perform the usual arithmetic operations with complex numbers in a natural way, e.g. $(a + ib)(c + id) = ac + iad + ibc + i^2bd = (ac - bd) + i(ad + bc)$. To every complex number $z = a + ib$, we can associate a point $X = (a, b)$ on the coordinate plane, and define the absolute value of z as the distance from X to the coordinate centre $O = (0, 0)$, that is, $|a + ib| = \sqrt{a^2 + b^2}$. The set $B_r(z_0) := \{z \in \mathbb{C} : |z - z_0| \leq r\}$ is called the *ball* with centre $z_0 \in \mathbb{C}$ and radius $r \in \mathbb{R}$. For every $z_1 = a + ib$ and $z_2 = c + id$,

$$|z_1 z_2| = |(a + ib)(c + id)| = \sqrt{(ac - bd)^2 + (ad + bc)^2} = \sqrt{a^2 + b^2} \sqrt{c^2 + d^2} = |z_1| |z_2|.$$

For every $f : \mathbb{C} \rightarrow \mathbb{C}$, let $K_f \subset \mathbb{C}$ be the set of initial points $z_0 \in \mathbb{C}$ such that the sequence

$$z_0, z_1 = f(z_0), z_2 = f(z_1), \dots, z_{n+1} = f(z_n), \dots \quad (4.1)$$

remains bounded, that is, $|z_n| \leq A$, $\forall n$ for some $A \in \mathbb{R}$. For example, for $f(z) = z^2$ we have $|z_n| = |z^{2^n}| = |z|^{2^n}$, so that the sequence is bounded if and only if $|z_0| \leq 1$. Geometrically, this means that $K_f = B_1(0)$ is the ball in the coordinate plane with centre $(0, 0)$ and radius 1. If f is the polynomial, the boundary of K_f is denoted by J_f and called the *Julia set* of f . By the definition of “boundary”, for every $z \in J_f$ and every $\varepsilon > 0$ there are numbers $z_0 \in K_f$ and $z'_0 \notin K_f$ such that $|z - z_0| < \varepsilon$ and $|z - z'_0| < \varepsilon$. In other words, the Julia set is the set of points for which the behaviour of the sequence (4.1) is extremely unstable: a very small perturbation of the initial point can cause drastic changes in the sequence, an unbounded one may become bounded and vice versa.

Julia Sets of Quadratic Polynomials

For $f(z) = z^2$, $K_f = \{z \mid |z| \leq 1\}$, and therefore its boundary is $J_f = \{z \mid |z| = 1\}$, that is, the Julia set is just a circle with centre $(0, 0)$ and radius 1. However, for slightly more complicated polynomials, such as $f(z) = z^2 + c$ or $f(z) = z^2 + cz$ for some $c \in \mathbb{C}$, the Julia set, for many values of c , becomes surprisingly complicated, see Fig. 4.1. For some c , the Julia set of $z^2 + c$ looks like a complex variant of a Cantor set, and there are even values of c such that J_{z^2+c} has (Hausdorff) dimension¹ 2! Recall that J_f is the boundary of K_f , which is a subset of the 2-dimensional complex plane \mathbb{C} . Can you imagine a set in the plane such that its boundary is 2-dimensional?

Given the complicated structure of a “typical” Julia set of a quadratic polynomial, any general result establishing its properties is very important. Petersen and Zakeri

¹See Sect. 3.7 for a formal definition of the Hausdorff dimension of any set.

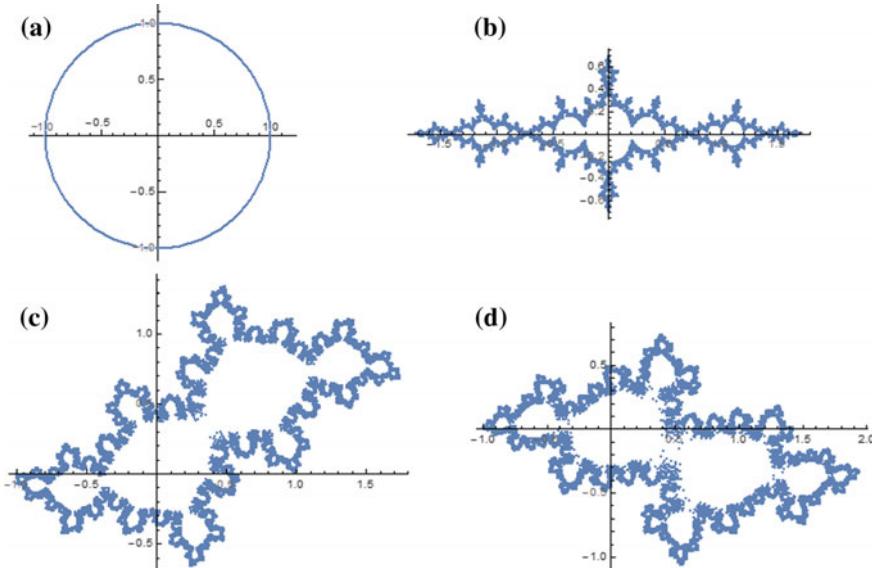


Fig. 4.1 Julia sets for **a** $f(z) = z^2$, **b** $f(z) = z^2 - 1.25$, **c** $f(z) = z^2 + e^{2\pi i \theta} z$ with $\theta = \frac{\sqrt{5}-1}{2}$, **d** $f(z) = z^2 + e^{2\pi i \theta} z$ with $\theta = [0; a_1, a_2, a_3, \dots]$, where $a_n = [e^{\sqrt{n}}]$

[302] proved a result of this kind: it states that for a “typical” quadratic polynomial f the Julia set J_f has Lebesgue measure 0 and is “locally connected”. Below we explain the meaning of these words.

Locally Connected Sets

A subset S of a 2-dimensional plane has Lebesgue measure 0 if it can be covered by a set of balls with total area ε for any $\varepsilon > 0$. For example, the unit disk $B_1(0) = \{z \mid |z| \leq 1\}$ has area π , and cannot be covered by any set of disks with total area ε for any $\varepsilon < \pi$. In contrast, the circle $C_1 = \{z \mid |z| = 1\}$ has length 2π and can be covered by N disks with centres lying uniformly along the circle and radius π/N each. The area of each disk is $\pi(\pi/N)^2 = \pi^2/N^2$, hence the total area is π^2/N . This total area can be made smaller than any given $\varepsilon > 0$ by selecting $N > \pi^2/\varepsilon$. Hence, C_1 has Lebesgue measure 0.

Further, a set S is called *open* if for every $z \in S$ there is an $\varepsilon > 0$ such that S contains $B_\varepsilon(z)$. For example, the disk $B_1(0) = \{z \mid |z| \leq 1\}$ is *not* an open set, because the condition in the definition does not work for any point $z \in B_1(0)$ with $|z| = 1$, for example, for $z = 1$. In contrast, the disk without boundary $B'_1(0) = \{z \mid |z| < 1\}$ is an open set, because the definition works with $\varepsilon = (1 - |z|)/2$.

A set S is called *connected* if it cannot be partitioned into two disjoint open sets, that is, there are no disjoint open sets V_1 and V_2 such that S is covered by

the union of V_1 and V_2 , and has a non-empty intersection with both V_1 and V_2 . For example, the circle $C_1 = \{z \mid |z| = 1\}$ is a connected set. However, the union of C_1 with $C_2 = \{z \mid |z| = 2\}$ (the notation for such a union is $C_1 \cup C_2$) is not connected, because, for example, $V_1 = \{z \mid |z| < \pi/3\}$ and $V_2 = \{z \mid |z| > \pi/3\}$ satisfy all the conditions in the above definition. The same choice of V_1 and V_2 shows that the set $R = \{z = x + iy \mid x \text{ and } y \text{ are rational}\}$ is not connected. Intuitively, however, these examples are very different if we look at them “locally”, in a small “neighbourhood” of any point. Around any $z \in C_1 \cup C_2$, the set $C_1 \cup C_2$ “looks like” a nice connected curve. In contrast, in any small area around any point $z \in R$, the set R is very far from “looking like” it is connected.

Formally, a set S is called *locally connected* if for every $z \in S$ and every open set U containing z there is an open set V containing z such that the intersection $V \cap S$ is connected and contained in U . These definitions may look technical, but the intuition is just that S “looks connected” locally. For example, $C_1 \cup C_2$ is a locally connected set, while R is not. In general, having measure 0 and being locally connected is what we would expect the boundary of a set to “naturally” look like.

Geometric Representation of Complex Numbers with Absolute Value One

Petersen and Zakeri studied quadratic polynomials of the form $f(z) = z^2 + cz$, where c is a complex number with $|c| = 1$. If $c = x + iy$, the point $A = (x, y)$ belongs to the unit circle $\{z \mid |z| = 1\}$ of the complex plane, and can be completely described by an angle ϕ between OA and the x -axis, that is, $x = \cos(\phi)$ and $y = \sin(\phi)$, $\phi \in [0, 2\pi]$. Using the parameter $\theta = \frac{\phi}{2\pi}$, we may write our family of quadratic polynomials as $f_\theta(z) = z^2 + (\cos(2\pi\theta) + i \sin(2\pi\theta))z$, where $\theta \in [0, 1]$. In fact, Euler’s famous formula states that $\cos(t) + i \sin(t) = e^{it}$ for all t , hence the expression for $f_\theta(z)$ simplifies to $f_\theta(z) = z^2 + e^{2\pi i \theta}z$.

Continued Fraction Expansion

Assume we want to approximate an irrational number, say π , by rational numbers. First, we can take the integer part, and write $\pi = 3 + \frac{1}{x_1}$ for some $x_1 > 1$. Ignoring $\frac{1}{x_1}$, we have an integer approximation $\pi \approx 3$. To get a better one, we can write $x_1 = 7 + \frac{1}{x_2}$, $x_2 > 1$, and ignore $\frac{1}{x_2}$ to get

$$\pi = 3 + \frac{1}{7 + \frac{1}{x_2}} \approx 3 + \frac{1}{7} = \frac{22}{7}.$$

Continuing in this way,

$$\pi = 3 + \frac{1}{7 + \frac{1}{15 + \frac{1}{\sqrt{3}}}} \approx 3 + \frac{1}{7 + \frac{1}{15}} = \frac{333}{106},$$

$$\pi = 3 + \frac{1}{7 + \frac{1}{15 + \frac{1}{1 + \frac{1}{\sqrt{4}}}}} \approx 3 + \frac{1}{7 + \frac{1}{15 + \frac{1}{1}}} = \frac{355}{113},$$

and so on. This is called a *continued fraction expansion*. A short notation for it is $[a_0; a_1, a_2, \dots, a_n, \dots]$, where a_i is the integer part of x_i . For example, $\pi = [3; 7, 15, 1, \dots]$.

The Theorem

We have now introduced all preliminary concepts and are ready to formulate the theorem.

Theorem 4.1 *Let E denote the set of irrational numbers $\theta = [a_1, a_2, a_3, \dots]$ such that*

$$\ln a_n < C_\theta \sqrt{n}, \quad \forall n \tag{4.2}$$

holds with some constant C_θ depending on θ . Then, for every $\theta \in E$, the Julia set J_f of $f(z) = z^2 + e^{2\pi i \theta} z$, is locally connected and has Lebesgue measure 0.

Figure 4.1c shows the Julia set for $f(z) = z^2 + e^{2\pi i \theta} z$ with $\theta = \frac{\sqrt{5}-1}{2} = [0; 1, 1, 1, 1, \dots]$, while Fig. 4.1d shows it for $f(z) = z^2 + e^{2\pi i \theta} z$ with $\theta = [0; a_1, a_2, a_3, \dots]$, $a_n = [e^{\sqrt{n}}]$, where $[t]$ denotes the largest integer not exceeding t . In both cases, condition (4.2) in Theorem 4.1 holds with $C_\theta = 1$, hence the corresponding θ belongs to the set E . In fact, Petersen and Zakeri show that the set of numbers $\theta \in [0, 1)$ not belonging to E has measure 0, so, in this sense, Theorem 4.1 is applicable to *almost all* quadratic polynomials $f_\theta(z)$.

The complement of the Julia set J_f is denoted F_f and is called the *Fatou set* of f . If the Julia set is the set of initial points z_0 with “unstable” dynamics, the Fatou set is the set of points z_0 such that the sequence (4.1) behaves “similarly” under small perturbations of z_0 . In fact, the Fatou set F_f can be partitioned into “components” F_f^1, F_f^2, \dots such that the “behaviour” of the sequence (4.1) depends, to a large extent, only on the component to which z_0 belongs. There exists one special component of F_f such that for every z_0 belonging to this component, the dynamics of (4.1) is particularly simple, in fact similar to rotation by an irrational angle. This special component is called the *Siegel disk* of f . Theorem 4.1 implies that, for almost all $\theta \in [0, 1)$, the Siegel disk of $f_\theta(z)$ is a so-called “Jordan domain”, which means that it has a very simple geometric structure.

Reference

C.L. Petersen and S. Zakeri, On the Julia set of a typical quadratic polynomial with a Siegel disk. *Annals of Mathematics* **159**-1 (2004): 1–52.

4.2 The Regular Polygons have Minimal Logarithmic Capacity

How to Find Quiet Places to Build Houses?

Assume that there is a short street where you can build houses for four families. Each family would like to live quietly, as far away from others as possible. Where should we build the houses to maximize the “average happiness” of everyone?

We can model the street as the interval $[-1, 1]$, and then build houses at coordinates $-1 \leq x_1 \leq x_2 \leq x_3 \leq x_4 \leq 1$. However, it is not so obvious how to model “average happiness”. The naïve approach would be to just maximize the average distance between all neighbours, that is,

$$M(x_1, x_2, x_3, x_4) := \frac{1}{6}[(x_2 - x_1) + (x_3 - x_1) + (x_4 - x_1) + (x_3 - x_2) + (x_4 - x_2) + (x_4 - x_3)]$$

where the multiplier $\frac{1}{6}$ arises because there are 6 terms in the sum. Because $(x_3 - x_1) + (x_4 - x_3) = (x_4 - x_1)$ and $(x_2 - x_1) + (x_4 - x_2) = (x_4 - x_1)$, the expression for $M(x_1, x_2, x_3, x_4)$ simplifies to $M(x_1, x_2, x_3, x_4) = \frac{1}{6}[3(x_4 - x_1) + (x_3 - x_2)]$. However, $x_4 - x_1 \leq 2$ and $x_3 - x_2 \leq 2$, hence $M(x_1, x_2, x_3, x_4) \leq \frac{4}{3}$, and equality holds if and only if $x_1 = x_2 = -1$ and $x_3 = x_4 = 1$. Unfortunately, it is clear that this is not the best way to build the houses. Every family has two neighbours at distance 2, but one neighbour in the same position, at distance 0. The average distance $\frac{1}{3}(2 + 2 + 0) = \frac{4}{3}$ is the maximal possible, but this does not make them happy: one neighbour at distance 0 outweighs all the advantages created by the maximal distance from other neighbours.

When One Bad Outcome Cancels All Good Ones

There are other examples where one bad outcome cancels all good ones. For example, driving at a higher speed than allowed can bring you a bit of “happiness” every day because of time saving, but this is only until you have a car accident. As another example, if you invest into a financial market, you can make a bit of profit every day, but this does not matter if you lose *all* the money during a crisis.

Let us look at the last example in more detail. Assume that there is a risky financial instrument *A* which can double your money with probability $2/3$ but you can lose everything with probability $1/3$, and another instrument *B* which promises you just 10% profit, but for sure. If you invest capital *C* into *A*, you can get $2C$, $2C$, or 0 with equal chances, $4C/3$ on average. In contrast, with *B* you get $1.1C$ for sure. Because $4C/3 > 1.1C$, should we select *A*? After all, we have a good chance to be lucky and get $2C$. Then we can invest $2C$ into *A* again, and, if lucky, get $4C$. Then invest again, and, if unlucky, lose everything. In contrast, investing 3 times into *B* would return us $(1.1)^3 C$.

An Alternative Way to Compare Investment Strategies

This suggest a better way to compare investments. If instrument A , after investing C , returns a_1C, a_2C, \dots, a_kC with equal chances, then investing all the money into A k times returns about $a_1a_2\dots a_kC$ (of course I may be lucky or unlucky and get more or less, respectively, but let us use this rough estimate). This is the same as investing k times into the instruments returning $\sqrt[k]{a_1a_2\dots a_k}$ for sure. Hence, A is better than another instrument B returning b_1C, b_2C, \dots, b_kC with equal chances if and only $\sqrt[k]{a_1a_2\dots a_k} \geq \sqrt[k]{b_1b_2\dots b_k}$. The quantity $\sqrt[k]{a_1a_2\dots a_k}$ is called the *geometric mean* of the numbers a_1, a_2, \dots, a_k . So, in this example, the geometric mean is a much better “measure of happiness” than the usual arithmetic mean $\frac{1}{k} \sum_{i=1}^k a_i$.

Back to Building Houses on a Street

Let us return to our example of building houses and try to maximize the *geometric mean* of all the distances

$$p(x_1, x_2, x_3, x_4) = \sqrt[4]{(x_2 - x_1)(x_3 - x_1)(x_4 - x_1)(x_3 - x_2)(x_4 - x_2)(x_4 - x_3)}.$$

It is intuitively obvious that, at optimality, $x_1 = -1$, $x_4 = 1$, and, from symmetry, $x_2 = -x_3$. Then

$$p(-1, -x_3, x_3, 1)^6 = 2x_3(1 - x_3)^2(1 + x^3)^2 = x_3^5 - 2x_3^3 + x_3,$$

from which it is not difficult to find that the optimal x_3 is $\sqrt[3]{\frac{1}{5}} \approx 0.45$. The houses at coordinates about $-1, -0.45, 0.45, 1$, look much more natural than the result of the arithmetic mean maximization.

What about optimal locations for $n \neq 4$ houses? When we build $n = 2$ houses on $[-1, 1]$, it is obvious that $p(x_1, x_2) = x_2 - x_1$ is maximal when $x_1 = -1$ and $x_2 = 1$. For $n = 3$, one can check that $p(x_1, x_2, x_3) = \sqrt[3]{(x_2 - x_1)(x_3 - x_1)(x_3 - x_2)}$ is maximal if $x_1 = -1$, $x_2 = 0$, $x_3 = 1$. Optimal locations for n houses, $2 \leq n \leq 6$, are presented in Fig. 4.2.

Building Houses in a City

Obviously, houses are usually built not along one street, but in a city, which can be modelled as a closed bounded subset E of the 2-dimensional plane \mathbb{R}^2 . The positions of k houses are points A_1, \dots, A_k belonging to E with coordinates $A_i = (x_i, y_i)$, $i = 1, \dots, k$. The distances between houses i and j are $|A_i - A_j| := \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. There are $\frac{k(k-1)}{2}$ such distances, and their geometric mean is

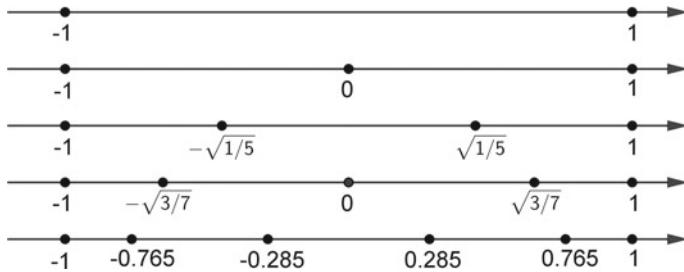


Fig. 4.2 Optimal locations for n houses, $2 \leq n \leq 6$

$$p(E, A_1, A_2, \dots, A_k) := \left(\prod_{i=1}^{k-1} \prod_{j=i+1}^k |A_j - A_i| \right)^{\frac{2}{k(k-1)}},$$

and the maximal possible “happiness” for k families is

$$p_k(E) := \max_{A_1, A_2, \dots, A_k} p(E, A_1, A_2, \dots, A_k).$$

For large cities, it makes sense to consider the limit

$$p(E) := \lim_{k \rightarrow \infty} p_k(E).$$

It is intuitively obvious and not difficult to formally prove that the more houses the less “alone” they can be in the same city, so $p_k(E) \geq p_{k+1}(E)$ for all $k \geq 2$. This together with the non-negativity of $p_k(E)$ implies that the limit always exists.

In particular, if E is a circle with radius 1, then one can show that houses should be distributed uniformly on the circle, and $p_k(E) = k^{\frac{1}{k-1}}$, resulting in $p(E) = 1$. If E is a line segment, as in the first example of this section, then $p(E) = h/4$, where h is the length of the segment.

The Logarithmic Capacity of a Plane Region

We have defined a way to assign to every closed bounded $E \subset \mathbb{R}^2$ a non-negative number $p(E)$. This number arises independently in surprisingly many ways, and therefore has a number of different names. With the definition as above it is usually called the *transfinite diameter* of E . Because the logarithm of a product of numbers is the sum of their logarithms, the same number arises from maximizing the arithmetic mean of logarithms of distances, and is therefore also called the *logarithmic capacity* of E . The number $p(E)$ also arises in the theory of polynomial approximation, and is named the *Chebyshev constant*. If a concept arises in many different natural ways and has several equivalent definitions, this is an indication of its importance.

The Minimal Logarithmic Capacity of a Polygon

In our example, $p(E)$ was used to model the level of “happiness” experienced by the residents of a city with a large number of households spread over an area A . Thus, it would be natural to ask what is the *minimal level of happiness* that can be guaranteed to the inhabitants of this city. It is also natural to assume that the city is squeezed between relatively straight portions of highways, rivers, mountain ranges, etc. and thus the city area is bounded by a polygon consisting of $n \geq 3$ straight line segments.

Under these assumptions, our question about the guaranteed minimal level of happiness can be reinterpreted as a question about the minimal logarithmic capacity among all closed polygons of area A bounded by n sides.

In 1951, Pólya and Szegő [307] proved that the equilateral triangle has minimal logarithmic capacity among all triangles with the same area, similarly the square among all quadrilaterals, and conjectured that the same is true for pentagons, hexagons and so on. It took more than 50 years until the conjecture was finally proved by Solynin and Zalgaller [352].

Theorem 4.2 *The regular closed polygon has the minimal logarithmic capacity among all closed polygons with the same number of sides and area.*

There is one more way to interpret the logarithmic capacity – as the maximal cost to build a network between a large number of households in a city if the cost is proportional to the average distance between households measured on the logarithmic scale. Then the shape that provides the minimal logarithmic capacity will also provide the minimal cost of the network.

Reference

A. Solynin and V. Zalgaller, An isoperimetric inequality for logarithmic capacity of polygons. *Annals of Mathematics* **159**-1 (2004): 277–303.

4.3 On the Growth of the Diffusion Coefficient

A Toy Model for the Diffusion Process

When you put a bit of milk into a cup of tea, it is initially concentrated in one region, but, after some time, the milk spreads out in the tea, finally mixing with it more or less uniformly. This process is called *diffusion*, and is the source of many important and difficult questions, for example, how fast should we expect the full mixing to happen?

When facing a difficult problem, mathematicians often try to build the simplest mathematical model to describe it, solve the problem within this model, and then try to make the model more complicated and realistic, and extend the solution of

the toy model to more general cases. So, let us build the simplest possible model which describes the diffusion of particles in a space. Let us start with the case when we have just one particle, moving along the 1-dimensional line. To simplify it even more, assume that the (unique) coordinate describing the position of the particle is an integer. At time $t = 0$ the particle starts at 0; then, at some random time t_1 it moves to the adjacent position -1 or 1 ; then, at some random time $t_2 > t_1$ it moves to the adjacent position again (for example, from 1 it can move to 0 or 2), and so on, see Fig. 4.3a.

Times of Movements and the Exponential Distribution

To finish the model, we need to specify exactly what we mean by “at some random time t_1 it moves...”. To start, let us first assume that particle can move only at some discrete moments of time, say, after $\frac{1}{n}$ of a second, or after $\frac{2}{n}$ of a second, and so on, where n is some large positive integer. At each of these moments, the particle moves with probability p and stays with probability $1 - p$. Then, after 1 second, there will be about pn jumps. Let us further assume that the “average speed” of the particle is 1 jump per second, hence $pn = 1$, and $p = 1/n$.

What is the chance that the particle will stay at the original position up to time $t = \frac{k}{n}$, where k is some integer? For this, it should stay at time $1/n$, stay at time $2/n$, and so on, up to time k/n . The probability of each “stay” is $1 - p = 1 - 1/n$. Hence, the chance that it stays all the time is $(1 - \frac{1}{n})^k = (1 - \frac{1}{n})^{nt}$. For large n , this expression can be approximated by its limiting value

$$P[\text{Stay up to time } t] = \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^{nt} = e^{-t},$$

where $e = \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n \approx 2.718$ is the base of the natural logarithm. However, “stay at the original position up to time t ” is exactly the same as “the time t_1 of the first move is greater than t ”, hence we have derived that $P[t_1 > t] = e^{-t}$, or, equivalently,

$$P[t_1 \leq t] = 1 - e^{-t}.$$

When this formula holds, we say that “the random variable t_1 follows the exponential distribution”. By a similar argument, the time $t_{k+1} - t_k$ between any two consecutive movements follows the exponential distribution.

Making the Toy Model More Realistic

Now we have a “diffusion model” for one particle on a line: it waits for some time according to the exponential distribution, then jumps to one of two adjacent positions, then waits again, jumps, and so on. The next step is to try to make this “toy” model

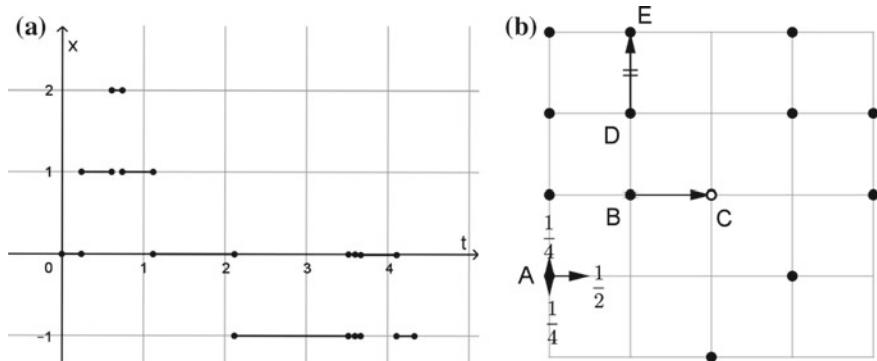


Fig. 4.3 **a** Evolution of a one-dimensional walk with time; **b** A two-dimensional asymmetric simple exclusion process

a bit more realistic. First, in practice we rarely study diffusion in a long thin tube, which can be modelled as a line, so we need an extension to higher dimensions. This is easy: for two dimensions, we can assume that the particle starts from the point $(0, 0)$ in the plane, waits for an exponential time, and then moves to one of the neighbours with integer coordinates: $(-1, 0)$, $(1, 0)$, $(0, 1)$, or $(0, -1)$; then waits again and moves the one of the neighbours again, and so on. The generalization to the three-dimensional space we live in, or even to arbitrary higher dimensions, works in exactly the same way.

Second, and this is the most important point, in reality we have *many* particles, and they *interact*. Let us initially put at each point (x_1, x_2) with integer coordinates either no particles or one particle, each variant with probability $\frac{1}{2}$, see Fig. 4.3b, and let them *all* move according to the rules described above. However, to model interaction, we assume that a particle can move to an adjacent site only if it is not occupied. If it is occupied, it stays where it was, waits exponential time, and tries to move again, and so on.

Finally, the reality may be *asymmetric*, and assuming that all four possible directions for movement are chosen with equal probability is not always an adequate model. We can define any rules of movement (for example, up, down, right, and left, with probabilities p_1, p_2, p_3, p_4 , respectively, for any non-negative p_i summing to 1, or even allow moves to some pre-specified non-adjacent positions), but, for a start, we define the following model: while movements up and down remain equally probable (symmetric), let us assume that left-right movements are totally asymmetric, that is, only jumps to the right are allowed. The model we have just defined is an example of a so-called *two-dimensional asymmetric simple exclusion process*.

Figure 4.3b illustrates the “state” of this process at some time t^* . The figure shows a fragment of the infinite plane containing 25 points with integer coefficients. 13 of these points happened to be “occupied” by a particle, and 12 happened to be “free”. Each particle waits exponential time, and then tries to move right, up, or down, with probabilities $1/2$, $1/4$, and $1/4$, respectively. At some moment $t_1 < t^*$, the particle

from position B happened to move to position C . Then, at some moment $t_2 < t^*$, the particle at position D tried to move to position E , but it was occupied, hence D stayed where it was. No other particles happened to move before time t^* .

Independence and Correlation

For any point $x = (x_1, x_2)$ with integer coordinates, let $\eta_x(t)$ be the number of particles (which can be 0 or 1) in position x at time t . For $t = 0$, all $\eta_x(0)$ are random variables, taking values 0 and 1 with equal chances, and they are all *independent*, because no interaction has happened yet. In particular, the independence implies that $E[\eta_x(0) \cdot \eta_y(0)] = E[\eta_x(0)] \cdot E[\eta_y(0)]$ for any different points $x = (x_1, x_2)$ and $y = (y_1, y_2)$, where E denotes the expectation, or average value of a random variable. Indeed, the average value of $\eta_x(0)$ is $E[\eta_x(0)] = \frac{0+1}{2} = 0.5$, and similarly $E[\eta_y(0)] = 0.5$. The product $\eta_x(0) \cdot \eta_y(0)$ can be $0 \cdot 0 = 0$, $0 \cdot 1 = 0$, $1 \cdot 0 = 0$ or $1 \cdot 1 = 1$ with equal chances, hence $E[\eta_x(0) \cdot \eta_y(0)] = 0.25 = 0.5 \cdot 0.5$.

However, if $t > 0$, the particle from $(0, 0)$ has time to move to x , and therefore the information that $\eta_{(0,0)}(0) = 1$ makes the chance $\eta_x(t) = 1$ greater than $\eta_x(t) = 0$. Hence $P[\eta_{(0,0)}(0) = \eta_x(t) = 1] > 0.25$, resulting in $E[\eta_x(t) \cdot \eta_{(0,0)}(0)] > 0.25 = E[\eta_x(t)] \cdot E[\eta_{(0,0)}(0)]$. The difference

$$S(x, t) := E[\eta_x(t) \cdot \eta_{(0,0)}(0)] - E[\eta_x(t)] \cdot E[\eta_{(0,0)}(0)] = E[\eta_x(t) \cdot \eta_{(0,0)}(0)] - 0.25,$$

can be used to measure how strong the correlation between random variables $\eta_{(0,0)}(0)$ and $\eta_x(t)$ is.

The Diffusion Coefficient

If $t = 0$, $S(x, 0) = 0.25 - 0.25 = 0$ for $x \neq (0, 0)$, but $S(x, 0) = 0.5 - 0.25 = 0.25$ for $x = (0, 0)$, hence $\sum_x S(x, 0) = 0.25$, where by \sum_x we mean $\sum_{x_1=-\infty}^{\infty} \sum_{x_2=-\infty}^{\infty}$.

Interestingly, the same relation holds for every t : $\sum_x S(x, t) = 0.25$, $\forall t \geq 0$. However, if for $t = 0$ all 0.25 of the correlation is “concentrated” at $(0, 0)$, for $t > 0$ the function $S(x, t)$ is “spread out”. To measure “how far” it spread out by the time t in the direction of the first coordinate, we can calculate the sum $\sum_x x_1^2 S(x, t) / \sum_x S(x, t) = \frac{1}{0.25} \sum_x x_1^2 S(x, t)$. Now, to measure the “speed” of this process, we divide this “distance” by time, and finally define

$$D_{11}(t) = \frac{4}{t} \sum_x x_1^2 S(x, t) = \frac{4}{t} \sum_{x_1=-\infty}^{\infty} \sum_{x_2=-\infty}^{\infty} x_1^2 S((x_1, x_2), t).$$

$D_{11}(t)$ is called the *diffusion coefficient* in direction 1. The following theorem of Yau [405] describes how fast it grows as a function of t .

Theorem 4.3 Consider the asymmetric simple exclusion process in dimension $d = 2$ as defined above. Then there exists a constant $\gamma > 0$ such that, for sufficiently small $\lambda > 0$,

$$\lambda^{-2} |\ln \lambda|^{2/3} e^{-\gamma |\ln \ln \ln \lambda|^2} \leq \int_0^\infty e^{-\lambda t} t D_{11}(t) dt \leq \lambda^{-2} |\ln \lambda|^{2/3} e^{\gamma |\ln \ln \ln \lambda|^2}.$$

The complicated formula in Theorem 4.3 is just the formalization (and refinement) of the simple fact that the diffusion coefficient $D_{11}(t)$ grows approximately as $(\ln t)^{2/3}$. According to Yau, the same technique may be applied to prove similar results in a wide class of related models.

Reference

H.-T. Yau, $(\ln t)^{2/3}$ law of the two dimensional asymmetric simple exclusion process. *Annals of Mathematics* **159**-1 (2004): 377–405.

4.4 On the Volume of the Intersection of Two Wiener Sausages

A Simple Model for Particle Movement

Can we describe the trajectory of a particle of liquid or gas in 3-dimensional space? Will it visit every point of the space if we wait sufficiently long? Will two given particles ever meet, or at least visit the same point in the space? To address this type of question we need first to build a model which describes particle movement. Because particle trajectories are very unpredictable, the best models are stochastic, i.e. those which allow particles to move randomly. First, imagine that a particle is moving in one dimension, along the line, can visit only discrete points on the line (say, with integer coordinates), and the movement can happen at discrete times only. The particle starts at time $t = 0$ at point 0, and then at times $t = 1, 2, 3, \dots$ moves one step right or left, with equal probabilities. For example, possible trajectories after 2 steps are $0 \rightarrow -1 \rightarrow -2$, $0 \rightarrow -1 \rightarrow 0$, $0 \rightarrow 1 \rightarrow 0$, or $0 \rightarrow 1 \rightarrow 2$, so the particle can move to -2 or 2 with probabilities 0.25, or return to 0 with probability 0.5.

Expectation and Variance

Let X_t be the position of the particle after t steps. Obviously, X_t can take several positions a_1, a_2, \dots, a_n with some probabilities p_1, p_2, \dots, p_n , and is therefore called a (discrete) random variable. Fundamental characteristics of any random variables are its expectation and variance. The expectation is just the average, that is, $E[X_t] =$

$\sum_{i=1}^n a_i p_i$; for example, $E[X_2] = (-2) \cdot 0.25 + (0) \cdot 0.5 + (+2) \cdot 0.25 = 0$. However, knowing the average $E[X_t]$ provides little information about X_t , unless we also know how close it is to the average, that is, the difference $X_t - E[X_t]$. The average $E[(X_t - E[X_t])^2]$ of the square of this distance is denoted by $\text{Var}[X_t]$ and called the *variance* of X_t .

Let Y_i be a random variable describing step i , that is, $Y_i = 1$ if the particle has moved right, and $Y_i = -1$ if it has moved left. Because X_t is the position after t such steps, $X_t = \sum_{i=1}^t Y_i$. It is known that expectation is a linear function, that is, $E[X_t] = \sum_{i=1}^t E[Y_i]$, and, because all steps Y_i are independent from each other, also $\text{Var}[X_t] = \sum_{i=1}^t \text{Var}[Y_i]$. But, for every i , $E[Y_i] = (-1) \cdot 0.5 + (+1) \cdot 0.5 = 0$, and $\text{Var}[Y_i] = E[(Y_i - E[Y_i])^2] = E[(Y_i - 0)^2] = (-1)^2 \cdot 0.5 + (+1)^2 \cdot 0.5 = 1$, hence $E[X_t] = 0$ and $\text{Var}[X_t] = \sum_{i=1}^t 1 = t$. We can conclude that, as t grows, the “average position” of the particle stays at 0, but the average distance from X_t to 0 grows with t , hence a particle will “typically” walk further and further from 0, and, in particular, will eventually visit every integer point on the line.

Points Visited and Unvisited

Of course, real particles move in 3-dimensional space, so let us extend this model to higher dimensions. Let \mathbb{Z}^d be the set of points $z = (z_1, z_2, \dots, z_d)$ in d -dimensional space such that all z_i , $i = 1, \dots, d$, are integers. Assume that a particle starts at $O = (0, 0, \dots, 0)$ and, at each step i , moves to any of the “adjacent” points with equal chance. For example, for $d = 2$ there are four possible moves, up and down and left and right, so position X_t after t steps is $X_t = \sum_{i=1}^t Y_i$, where each Y_i takes values $(0, 1), (0, -1), (-1, 0), (1, 0)$ with equal chances. For general d , Y_i can take $2d$ values in the form $(0, \dots, 0, \pm 1, 0, \dots, 0)$, each with probability $1/2d$. In this case, we can also prove that $\text{Var}[X_t]$ grows with t , and the particle will “typically” walk further and further from O . However, while for $d = 1$ it must eventually visit every point, being unable to go “around”, say, point 5, for $d = 3$ it could easily move away from O without visiting, say, the point $(3, 5, 7)$. In fact, Pólya’s Theorem implies that, if $d \geq 3$, then, with probability 1, there are points which the particle will never visit, no matter how long we wait! In other words, the set $R = \{z \in \mathbb{Z}^d : X_t = z \text{ for some } t \in \mathbb{N}\}$ of points which it will eventually visit is a proper subset of \mathbb{Z}^d .

Given this, the following question naturally arises: if we have two particles moving independently, and $R_1, R_2 \subset \mathbb{Z}^d$ are the sets of points they will eventually visit, what can we say about the set $R_1 \cap R_2$ that *both* particles will visit? It turns out that, if $d \geq 5$, then, with probability 1, the number $|R_1 \cap R_2|$ of elements in this set is finite! That is, there are only a finite number of points that the particles will both visit, and then move off in different directions, never to intersect again, no matter how long we wait. Luckily, we do not live in 5-dimensional space, and for $d \leq 4$ $|R_1 \cap R_2| = \infty$ with probability 1. So, in our 3-dimensional space we can observe as many intersections as we like, if we wait long enough. The question is how long

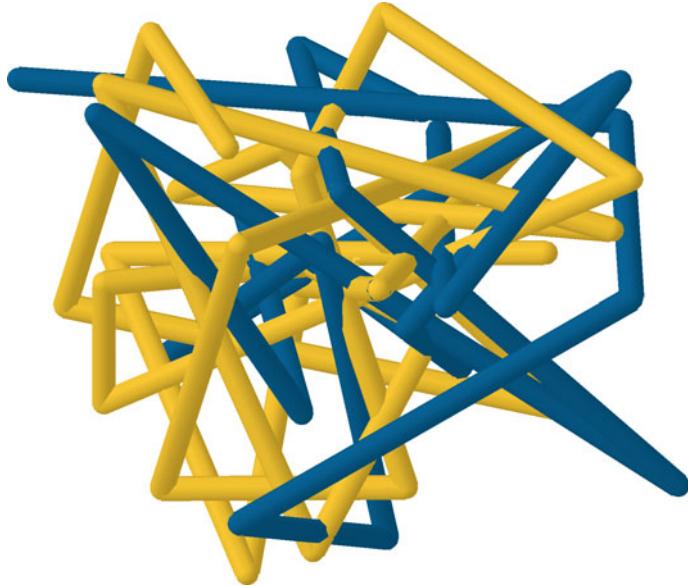


Fig. 4.4 Intersection of two Wiener sausages

we should wait. But, before we consider it, let us modify our model and make it more realistic. After all, real particle trajectories are continuous, not discrete.

The Wiener Process and Wiener Sausages

Let us return to dimension $d = 1$. In our model, $X_t = \sum_{i=1}^t Y_i$, where each step Y_i is either -1 or 1 . However, a real particle would not move in t big steps, its trajectory would rather look like a huge amount N of very small steps. So, let us write $X_t = \sum_{i=1}^N Z_i$, where each step Z_i is either $-\varepsilon$ or ε for some small $\varepsilon > 0$. Then $\text{Var}[Z_i] = \varepsilon^2$, and $\text{Var}[X_t] = \sum_{i=1}^N \text{Var}[Z_i] = N\varepsilon^2$. However, we know that $\text{Var}[X_t] = t$, so in fact $t = N\varepsilon^2$, so we should select $\varepsilon = \sqrt{t/N}$. Now, the limit

$$W(t) = \lim_{N \rightarrow \infty} \sum_{i=1}^N Z_i, \quad \forall t > 0,$$

is a good model for the one-dimensional movement of a particle, and is called Brownian motion, or the Wiener process. If $d = 1$, the trajectory of a particle will cover the whole line with probability 1, and the trajectories of any two particles intersect infinitely often.

For $d \geq 2$, the trajectories of any two particles modelled as a point would never intersect, just because they are infinitely thin. The correct model for particle trajectory in this case is to assume that a particle is a ball with small but positive radius $a > 0$ and define

$$W^a(t) = \{x \in \mathbb{R}^d \mid \exists s \in [0, t] : \rho(x, W(s)) \leq a\},$$

where ρ is the usual distance in \mathbb{R}^d . The trajectory $W^a(t)$ up to time t looks like a “sausage” of radius a whose centreline is Wiener process $W(s)$, and is called a *Wiener sausage*, see Sect. 1.1.

The Volume of the Intersection of Two Wiener Sausages

Now, if we have two particles moving independently with trajectories $W_1^a(t)$ and $W_2^a(t)$, what can we say about the intersection

$$V^a(t) = W_1^a(t) \cap W_2^a(t),$$

depicted in Fig. 4.4? This is the subject of the following Theorem of M. van den Berg, E. Bolthausen and F. den Hollander [385].

Theorem 4.4 *For every $a > 0, c > 0$,*

$$\lim_{t \rightarrow \infty} \frac{1}{\ln t} \ln P(|V^a(ct)| \geq t/\ln t) = -I_2^{2\pi}(c), \quad d = 2,$$

and

$$\lim_{t \rightarrow \infty} \frac{1}{t^{(d-2)/d}} \ln P(|V^a(ct)| \geq t) = -I_d^{C(a,d)}(c), \quad d \geq 3.$$

Here, $I_d^{C(a,d)}(c)$ are the “rate constants”, for which explicit formulas are presented, and $C(a, d)$ will be defined in a moment. The average volume $E|W^a(t)|$ of the Wiener sausage $W^a(t)$ is equal to $2\pi t / \ln t$ if $d = 2$, and $C(a, d)t$ for $d \geq 3$, where $C(a, d)$ is a constant depending on a and d , see Sect. 1.1. Hence, Theorem 4.4 estimates the probability that $|V^a(t)|$ will differ from the average values of $|W_1^a(t)|$ and $|W_2^a(t)|$ by at most a constant factor.

Reference

M. van den Berg, E. Bolthausen and F. den Hollander, On the volume of the intersection of two Wiener sausages, *Annals of Mathematics* **159**-2, (2004), 741–783.

4.5 Controlling the Size of the Bilinear Hilbert Transform

Several Different Criteria for Comparing Errors

Assume that you have two methods to predict tomorrow's temperature outside, or tomorrow's dollar to pound exchange rate. How would you compare which one is better? The most obvious approach is to apply both, and let them return results a and b . Then you wait until tomorrow to observe the actual value c of the quantity they predicted, and conclude that their errors are $x = a - c$, $y = b - c$, respectively. Finally, you can choose the method for which the absolute value of the error is lower.

However, what if this method was just lucky? To make an informed choice, it is better to repeat the experiments n times, observing errors $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$. Then you can choose the method for which the *average* absolute value of error is lower, that is, choose x if and only if $\frac{1}{n} \sum_{i=1}^n |x_i| \leq \frac{1}{n} \sum_{i=1}^n |y_i|$. Obviously, the multiplier $\frac{1}{n}$ cancels out, and we can equivalently compare the *total* error, that is, choose x if $A_1(x) \leq A_1(y)$ where $A_1(z) = \sum_{i=1}^n |z_i|$ for any $z = (z_1, z_2, \dots, z_n)$. However, if for $n = 1$ it was clear that we should compare absolute values of errors, for general n the choice of $A_1(z)$ to judge the quality of the method with error z is questionable. For example, if we represent z as a point in n -dimensional space, then its usual Euclidean norm is $A_2(z) = \sqrt{\sum_{i=1}^n z_i^2}$, so why not choose the method whose error has lower Euclidean norm in \mathbb{R}^n ? For example, if $n = 3$, $x = (5, 0, 0)$, $y = (2, 2, 2)$, then $A_1(x) = 5 < 6 = A_1(y)$ but $A_2(x) = \sqrt{25} > \sqrt{12} = A_2(y)$, so, compared to the A_1 criterion, the A_2 criterion penalizes methods which may work better most of the time, but sometimes produce very large errors. If we want to penalize such methods even more, we may select the criterion $A_3(z) = \sqrt[3]{\sum_{i=1}^n z_i^3}$, or, more generally, $A_p(z) = \sqrt[p]{\sum_{i=1}^n z_i^p}$ for any $p \in [1, \infty)$, the higher p the higher penalties for large errors. In the extreme case, we can just select a method whose highest error is lower, that is, choose x if and only if $A_\infty(x) \leq A_\infty(y)$, where $A_\infty(z) = \max_{i=1,\dots,n} |z_i|$.

From Discrete to Continuous: Comparing Functions

More generally, what if we need a continuous prediction, like the temperature outside or the dollar to pound exchange rate, as a function of time? Let $a(t)$ and $b(t)$ be the predictions from two methods, $c(t)$ be the actual value we later observe, and $x(t) = a(t) - c(t)$ and $y(t) = b(t) - c(t)$ be the respective errors. Now we should compare not numbers or n -dimensional vectors, but functions. How do we decide that one function is “smaller” than another one? To estimate the “average” or “total” of a function, we need to replace summation by integration, see e.g. Sect. 3.3 to recall the definitions of integral and derivative. So, we can say that a function $x : \mathbb{R} \rightarrow \mathbb{R}$ is “smaller” than $y : \mathbb{R} \rightarrow \mathbb{R}$ if and only if $\|x\|_1 \leq \|y\|_1$, where $\|z\|_1 = \int_{-\infty}^{\infty} |z(t)| dt$. If we would like to penalize methods producing large errors, we can also make a

decision on the basis of the criterion

$$\|z\|_p = \left(\int_{-\infty}^{\infty} |z(t)|^p dt \right)^{1/p}$$

for any $p \in [1, \infty)$.

Conditions for Fast Decrease: Schwartz Functions

Of course, for some functions, like $z(t) = t$, the integral above is infinite, and $\|z\|_p = \infty$, so we need to impose some extra assumptions on z to ensure that $\|z\|_p$ exists and is finite. One assumption that surely suffices to guarantee the finiteness of $\|z\|_p$ for all p is to assume that z is a Schwartz function. Intuitively, these are the functions which are rapidly decreasing as $|t| \rightarrow \infty$, and so are all their derivatives. For every $k \geq 2$, the k -th derivative of a function $z(t)$, denoted $z^{(k)}(t)$, is the derivative of $z^{(k-1)}(t)$, where $z^{(k-1)}(t)$ is the derivative of $z^{(k-2)}(t)$, and so on, by induction, until we reach $z^{(1)}(t) := z'(t)$ which is the derivative of the original function $z^{(0)}(t) := z(t)$. Now, assume that for a function $z : \mathbb{R} \rightarrow \mathbb{R}$ there exist all derivatives $z^{(k)}(t)$ for all $k \in \mathbb{N}$ and all $t \in \mathbb{R}$, and, for every k and $\gamma \in \mathbb{R}$, there is a constant $C(k, \gamma)$ such that

$$|t^\gamma z^{(k)}(t)| \leq C(k, \gamma) \quad \forall t \in \mathbb{R}. \quad (4.3)$$

Then z is called a *Schwartz function*. The definition implies that $|z^{(k)}(t)| \leq C(k, \gamma)/t^\gamma$, so that $z^{(k)}(t)$ decreases really fast, faster than *any* power of t . For example, the function $f(t) = t^{-100}$ decreases pretty fast, but still is not a Schwartz function because (4.3) fails for $k = 0$ and $\gamma = 101$. Indeed, in this case (4.3) reduces to $|t^{101}t^{-100}| = |t| \leq C(0, 101)$, $\forall t \in \mathbb{R}$, which is obviously false for every constant $C(0, 101)$.

Schwartz Functions: Examples and Some Properties

Given this, it is not even obvious that Schwartz functions exist, but there are actually a lot of them. For example, for any $k > 0$ and $\alpha > 0$, the function $z(t) = t^k e^{-\alpha t^2}$ is a Schwartz function (Fig. 4.5). Also, assume that z is an infinitely differentiable function such that $z(t) = 0$ outside of some interval $[a, b]$. Then, obviously, any derivative $z^{(k)}(t)$ is also 0 outside of $[a, b]$. Because $z^{(k)}(t)$ is differentiable, it is continuous, hence $t^\gamma z^{(k)}(t)$ is also continuous, and therefore bounded on $[a, b]$ by some constant $C(k, \gamma)$.

For any Schwartz function z , and any $p \in [1, \infty)$, let us apply (4.3) with $k = 0$ and $\gamma = 2/p$. Then $|z(t)| \leq C(0, 2/p)/t^{2/p}$, hence $|z(t)|^p \leq C(0, 2/p)^p/t^2$, and

$$\int_1^\infty |z(t)|^p dt \leq C(0, 2/p)^p \int_1^\infty \frac{dt}{t^2} = C(0, 2/p)^p < \infty.$$

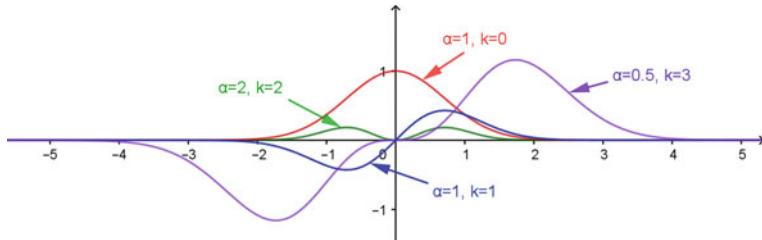


Fig. 4.5 Some examples of Schwartz functions from the family $z(t) = t^k e^{-\alpha t^2}$

We can prove $\int_{-\infty}^{-1} |z(t)|^p dt < \infty$ by the same argument, and $\int_{-1}^1 |z(t)|^p dt < \infty$ by continuity of $|z(t)|^p$. Hence $\|z\|_p < \infty$ is guaranteed for any Schwartz function z .

Controlling the “Size” of the Bilinear Hilbert Transform

In 1960, Calderón, [84] while trying to prove a difficult result, introduced a method which takes two functions $f : \mathbb{R} \rightarrow \mathbb{R}$ and $g : \mathbb{R} \rightarrow \mathbb{R}$, and returns another function $z_{\alpha,\beta}(x) : \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$z_{\alpha,\beta}(x) = \lim_{\varepsilon \rightarrow 0} \int_{\frac{1}{\varepsilon} > |t| > \varepsilon} f(x - \alpha t) g(x - \beta t) \frac{dt}{t}. \quad (4.4)$$

Here, $\alpha, \beta \in \mathbb{R}$ are real parameters. A transformation of the form (4.4) is called a *bilinear Hilbert transform*. If f and g are Schwartz functions, $z_{\alpha,\beta}$ is always well-defined. Calderón proved that his theorem would follow from an easy-looking lemma about $z_{\alpha,\beta}$, but he could not prove the lemma. His theorem was then proved by another method, but the “lemma” turned out to be more difficult and remained open for more than 40 years. People have found other applications of it, but could not prove it. In 2004, it was proved by Grafakos and Li [170] as a corollary² of the following theorem.

Theorem 4.5 *Let $2 < p_1 < \infty$, $2 < p_2 < \infty$, and $1 < p = \frac{p_1 p_2}{p_1 + p_2} < 2$. Then there is a constant $C = C(p_1, p_2)$ such that for all Schwartz functions $f, g : \mathbb{R} \rightarrow \mathbb{R}$,*

$$\sup_{\alpha, \beta} \|z_{\alpha,\beta}\|_p \leq C \|f\|_{p_1} \|g\|_{p_2}, \quad (4.5)$$

where $z_{\alpha,\beta}$ is given by (4.4).

Theorem 4.5 states that we have control of “how big” the function $z_{\alpha,\beta}$ given by (4.4) is, based on information about “how big” the functions f and g are. The

²To be precise, Calderón’s lemma follows from Theorem 4.5 in combination with the main result of another paper of Xiaochun Li.

difficulty and importance of Theorem 4.5 is that the constant in (4.5) is the same for all parameters α and β . In such cases we say that $z_{\alpha,\beta}$ is *uniformly* bounded with respect to the parameters α, β .

Reference

L. Grafakos and X. Li, Uniform bounds for the bilinear Hilbert transforms I, *Annals of Mathematics* **159**-3, (2004), 889–933.

4.6 Covering Convex Bodies by Balls

Tiling a Chess Board with Dominos

A fun old problem asks how many 1×2 dominos we need to cover an 8×8 chess board. Well, in this formulation this is easy – 32 clearly suffices, and 31 would cover a total area of only 62, while the area of the chess board is 64. However, the problem actually asks to cover a chess board in which two opposite corner squares (for example $a1$ and $h8$ in chess notation) have been removed. In this case the remaining area is 62, so, why not try to cover it with 31 dominos? If you are seeing this problem for the first time, you can try, and you will fail, and here is why. The squares on opposite corners of the chess board, coloured in a standard way, are of the same colour, say black, hence the remaining board has 32 white squares, with total area 32. However, any domino, no matter how we place it, covers at most 1 white square, so at least 32 dominos are required.

Distributing Police within a City

This was just fun, but very similar problems have numerous practical applications. For example, imagine you need a square of size 1×1 in a city to be controlled by police, and each policeman can see a distance of at most r , that is, control an area in the form of a disk of radius r around him. How many policemen do we need, say, for $r = 0.7$? In this case one policeman can cover the area $\pi r^2 > 1.5$, far greater than 1, however, he cannot control the whole square no matter where we place him. This is because any two points that can be controlled by a policeman are at distance at most $2r = 1.4$ from each other, while the non-adjacent vertices of the square are at distance $\sqrt{2} > 1.4$. On the other hand, two policemen clearly suffice. For example, if we choose the coordinate plane such that the vertices of the square are $A = (0, 0)$, $B = (1, 0)$, $C = (1, 1)$, and $D = (0, 1)$, then policemen at positions $(0.25, 0.5)$ and $(0.75, 0.5)$ can control the full square even if they can see a distance of $r = \frac{\sqrt{5}}{4} \approx 0.56 < 0.7$, see Fig. 4.6.

How many policemen do we need if, say, $r = 0.5$? Ok, 4 of them at coordinates $(0.25, 0.25)$, $(0.75, 0.25)$, $(0.75, 0.75)$, $(0.25, 0.75)$ suffices even for $r = \frac{1}{2\sqrt{2}} \approx 0.35 < 0.5$, but can we have 3? In this case there would be 2 vertices of the

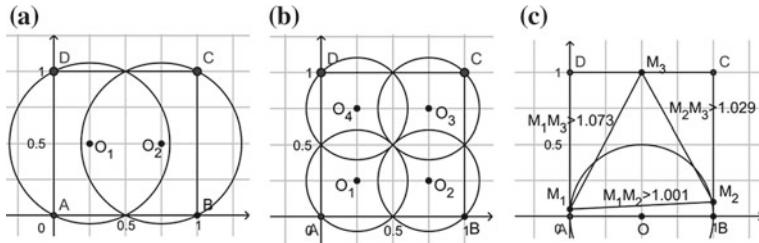


Fig. 4.6 Covering squares by disks

square covered by one policeman, and this is possible only if he were located at the midpoint of the corresponding side. Without loss of generality, let him be located at the midpoint O of side AB , whose coordinates are $(0.5, 0)$, see Fig. 4.6. Then points $M_1 = (0, 0.05)$, $M_2 = (1, 0.1)$, and $M_3 = (0.5, 1)$ are uncovered by him, the distance between any two of these points is greater than $2r = 1$ (for example, $\sqrt{(0.5 - 1)^2 + (1 - 0.1)^2} = \sqrt{1.06} > 1$), hence we need at least 3 more policemen to cover them, resulting in at least 4 policemen in total.

A More General Problem

Geometrically, this problem asks to cover a square by disks. Obviously, the question also makes sense for any other shapes, for example we can ask for a cover of a rectangle by disks, or a disk by squares, etc. The same problem can be formulated in any dimension, for example, how many balls do we need to cover a cube?

In general, n -dimensional Euclidean space \mathbb{R}^n is a set of points with coordinates (x_1, x_2, \dots, x_n) . A set $T \subset \mathbb{R}^n$ is called a *convex body* if it is convex, bounded, closed, and has non-empty interior, see Sect. 2.13 for more details about this definition. For any convex body T and $x \in \mathbb{R}^n$ we will write $x + T$ for a body T translated by a vector x , that is, $x + T = \{y \in \mathbb{R}^n, | y = x + t \text{ for some } t \in T\}$. For any two convex bodies $K \subset \mathbb{R}^n$ and $T \subset \mathbb{R}^n$ let

$$N(K, T) := \min \left\{ N : K \subset \bigcup_{i=1}^N (x_i + T) \text{ for some } x_1 \in \mathbb{R}^n, x_2 \in \mathbb{R}^n, \dots, x_N \in \mathbb{R}^n \right\}.$$

In other words, N is the minimal number of translates of T necessary to cover K . Note that we do not allow rotation, only translation. For example, we show above that $N(K, T) = 4$ if K is the unit square and T is the disk of radius $r = 0.5$, which are convex bodies in dimension $n = 2$. However, if in reality T is a model of a city, it is unrealistic to assume that one policeman can control half of its size, so the problem becomes more interesting if $r = 0.01$ or so. How do we estimate the number of policemen in this case? A careful case-by-case analysis as above can be

difficult for hundreds of policemen, and rough area-based estimates work extremely badly, because disks cannot cover any body in an area-efficient way: however we arrange them, they will heavily overlap, and, to cover all points in the area, we need to cover some multiple times.

Covering Squares by Disks and Disks by Squares

The situation would be much easier for the *opposite* problem: cover a large disk K by small squares T . In contrast to disks, squares can cover the full plane without any overlap, so the total number of squares we need can be well approximated by the ratio of areas K and T . Of course, there are boundary effects, but they become less and less important when $N(K, T)$ grows: after all, for large cities, only a tiny fraction of policemen will be placed anywhere close to the city boundary.

This is good, but looks totally unrelated to our original problem. How could covering a disk by squares help to cover a square by disks? It turns out there is a deep connection between the problems “how to cover a disk” and “how to cover by disks”, and this is what the theorem below is about.

Polar Bodies

We need some more definitions. We say that a convex body T is *symmetric* with respect to the origin if $x \in T$ implies that $-x \in T$. For any convex body T and $\alpha \in \mathbb{R}$ define $\alpha T := \{y \in \mathbb{R}^n, |y = \alpha t \text{ for some } t \in T\}$ to be the copy of T scaled by α . Finally, for a convex body T , define

$$T^\circ := \{u \in \mathbb{R}^n : \sup_{t \in T} \langle t, u \rangle \leq 1\},$$

where $\langle t, u \rangle = \sum_{i=1}^n t_i u_i$ is the scalar product in \mathbb{R}^n . T° is called the *polar body* of T . For example, the polar body of the unit disk is again the unit disk, and the same is true in any dimension. Indeed, let $T = \{t : \|t\| \leq 1\}$, where $\|\cdot\|$ is the usual norm in \mathbb{R}^n , be the unit ball. The well-known Cauchy–Schwarz inequality states that $\langle t, u \rangle \leq \|t\| \cdot \|u\|$, hence $\sup_{t \in T} \langle t, u \rangle \leq 1$ for every u such that $\|u\| \leq 1$. On the other hand, if $\|u\| > 1$, then $\sup_{t \in T} \langle t, u \rangle \geq \langle u, u \rangle = \|u\|^2 > 1$. This implies that $T^\circ := \{u : \|u\| \leq 1\}$, as promised. As another example, one can show that the polar body of the square with vertices $(\pm 1, \pm 1)$ is another square, just rotated and scaled.

The Connection Between “Covering by Balls” and “Covering a Ball”

We are ready to formulate the theorem of Artstein, Milman, and Szarek [17].

Theorem 4.6 *There exist two universal constants α and β such that for any dimension n and any convex body $K \subset \mathbb{R}^n$, symmetric with respect to the origin,*

$$N(D, \alpha^{-1} K^\circ)^{\frac{1}{\beta}} \leq N(K, D) \leq N(D, \alpha K^\circ)^\beta,$$

where K° is the polar body of K , and $D := \{u : \|u\| \leq 1\}$ is the Euclidean unit ball.

Because the polar body of a square is another square, Theorem 4.6, applied to the case when K is the (symmetric) square, relates the number $N(K, D)$ of squares needed to cover the disk to the number $N(D, \alpha K^\circ)$ of circles needed to cover a square of different size. Hence, the solution of the easier problem can be used as an estimate for the solution of the more difficult one. Of course, this is just a special case, the theorem works for any symmetric convex body K in any dimension, and the constants α and β in it are universal: they are the same for all bodies K and for all dimensions. Moreover, the authors show that we can select β as low as $2 + \varepsilon$ for any $\varepsilon > 0$. The theorem has application to, for example, “the duality conjecture for entropy numbers of linear operators”, posed by Pietsch [305] in 1972, which states that there exist two constants $a, b \geq 1$ such that

$$\ln N(T^\circ, aK^\circ) \leq b \ln N(K, T)$$

holds for any symmetric convex bodies K and T in any dimension n . Theorem 4.6 verifies this conjecture when either K or T are balls, or, more generally, ellipsoids.

Reference

S. Artstein, V. Milman, and S.J. Szarek, Duality of metric entropy, *Annals of Mathematics* **159**-3, (2004), 1313–1328.

4.7 The Parametrization of Quartic Rings

Groups and Their Isomorphisms

A *group* is a set G together with a binary operation $+$, called *addition*, such that (i) $(a + b) + c = a + (b + c)$ for all $a, b, c \in G$; (ii) there exists an element $0 \in G$ such that $a + 0 = 0 + a = a$ for all $a \in G$; and (iii) for every $a \in G$, there exists an element $-a \in G$, such that $a + (-a) = (-a) + a = 0$. A group G is called *abelian* if also (iv) $a + b = b + a$ for all $a, b \in G$. The simplest example of an infinite abelian group is the set of integers \mathbb{Z} with the usual addition. More generally, the set of n -dimensional vectors $x = (x_1, x_2, \dots, x_n)$ with integer coordinates $x_i \in \mathbb{Z}$, $i = 1, \dots, n$, forms an abelian group with natural component-wise addition $(x_1, x_2, \dots, x_n) + (y_1, y_2, \dots, y_n) = (x_1 + y_1, x_2 + y_2, \dots, x_n + y_n)$, $0 = (0, 0, \dots, 0)$, and $-(x_1, x_2, \dots, x_n) = (-x_1, -x_2, \dots, -x_n)$. Verifying that properties (i)–(iv) hold in this example is an easy exercise. This group is called \mathbb{Z}^n .

We say that two groups G and H are *isomorphic* if there exists a one-to-one correspondence between their elements which preserves the group operations. For example, let e_1, e_2, \dots, e_n be any set of linearly independent vectors in \mathbb{R}^n , possibly

with non-integer or even irrational coordinates. Let G be the set of vectors which can be represented in the form $x = \sum_{i=1}^n x_i e_i$, with $x_i \in \mathbb{Z}$, $i = 1, \dots, n$. Then G forms a group with a natural addition operation, and this group is isomorphic to \mathbb{Z}^n , because we can put each element of G as above into correspondence with an element $(x_1, x_2, \dots, x_n) \in \mathbb{Z}^n$.

Groups of Matrices with Integer Entries

A more interesting example of a group is the $n \times n$ matrices with integer entries, see Sect. 2.2 for the definition of matrices, their addition and multiplication. If our group operation is matrix addition, then this group is isomorphic to \mathbb{Z}^{n^2} , because, for addition, it makes no difference whether we write n^2 integers in the form of an $n \times n$ matrix or just an n^2 -long vector. However, let us choose matrix *multiplication* as the group operation. Then property (i) is still true, property (ii) holds if we choose the matrix E_n with all diagonal entries 1 and all non-diagonal entries 0 as the “0” element, but property (iii) states that, for a matrix A with integer entries, there exists a matrix B with integer entries such that $BA = AB = E_n$. This is not always true, but let us denote by $GL_n(\mathbb{Z})$ the set of those $n \times n$ matrices A with integer entries for which such B exists. Then $GL_n(\mathbb{Z})$ is a group with matrix multiplication as the operation. Note that property (iv) does not hold for this group, that is, $GL_n(\mathbb{Z})$ is *not* abelian.

Commutative Rings and Their Ranks

A *commutative ring* is a set R together with *two* binary operations $+$ and \cdot , called addition and multiplication, respectively, such that R with addition $+$ is an abelian group, that is, (i)–(iv) above hold, and also (v) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all $a, b, c \in R$; (vi) there exists an element $1 \in R$ such that $a \cdot 1 = 1 \cdot a = a$ for all $a \in R$; (vii) $a \cdot b = b \cdot a$ for all $a, b \in R$; (viii) $a \cdot (b + c) = a \cdot b + a \cdot c$ for all $a, b, c \in R$, where the last expression $a \cdot b + a \cdot c$ is understood with usual convention that we first perform multiplication and then addition. Of course, a classical example of a commutative ring is again the set of integers \mathbb{Z} . Similarly to groups, two rings R and Q are called *isomorphic* if there exists a one-to-one correspondence between their elements which preserves both operations, see Sect. 1.5 for more examples of groups and rings, isomorphic and not.

By definition, each ring is also an abelian group with respect to addition. A commutative ring is called a *ring of rank n* if the corresponding abelian group is isomorphic to \mathbb{Z}^n . By definition, the ring of integers \mathbb{Z} is the ring of rank 1, and one can easily show that this is *the only* ring of rank 1, that is, there is no way to define “multiplication” of integers, other than the standard one, such that axioms (v)–(viii) hold.

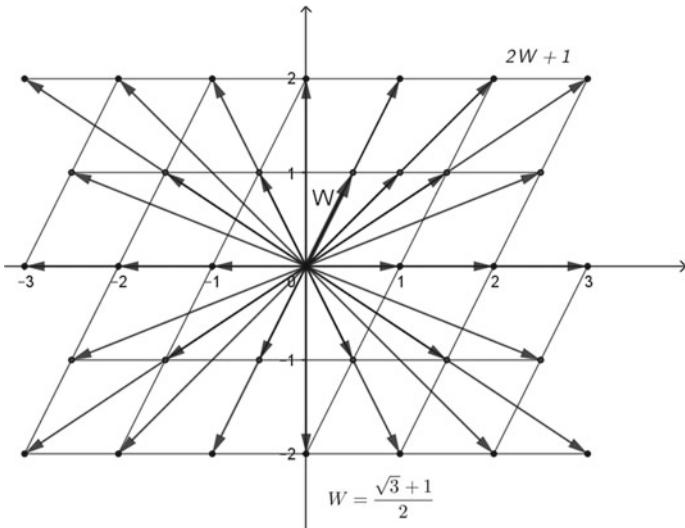


Fig. 4.7 An example of a quadratic ring

Classification of Quadratic Rings

The situation is more interesting for rings of rank 2, which are also called *quadratic rings*. Because they are isomorphic to \mathbb{Z}^2 with respect to addition, every element x of such a ring R can be written as $x = a \cdot e_1 + b \cdot e_2$ for some $e_1, e_2 \in R$ and $a, b \in \mathbb{Z}$. We can choose $e_1 = 1$, where 1 is defined in property (vi), and denote e_2 by w , so that x can be written as $x = a \cdot 1 + b \cdot w = a + bw$, where the \cdot operation in the expression $b \cdot w$ is omitted. If $y = c + dw$ is another element of R , then the properties in the definition of a ring imply that $xy = (a + bw)(c + dw) = ac + (ad + bc)w + bd(ww)$. So, to define a multiplication in R , we only need to define $w^2 = w \cdot w$, and different choices lead us to different rings. The simplest choice is $w^2 = 0$, then $xy = (a + bw)(c + dw) = ac + (ad + bc)w$, and this can be interpreted as a ring of polynomials in one variable w with integer coefficients, such that terms near w^k , $k \geq 2$, are ignored. The choice $w^2 = -1$ leads to the multiplication rule $xy = (a + bw)(c + dw) = (ac - bd) + (ad + bc)w$, and, with the notation $w = i$, this is exactly the ring of all complex numbers $a + bi$ with $a, b \in \mathbb{Z}$. More generally, assuming $w^2 = k \cdot 1$ for some integer k , we get the ring of all numbers in the form $a + b\sqrt{k}$, $a, b \in \mathbb{Z}$. Finally, we can also assume that $w^2 = w + k \cdot 1$ for some integer k . In this case, the multiplication rule becomes $(a + bw)(c + dw) = ac + (ad + bc)w + bd(w + k) = (ac + bdk) + (ad + bc + bd)w$. The simplest example of such a ring is the ring of all complex numbers representable as $a + bw$, where a, b are integers, and w is one of two roots of the equation $w^2 = w + k$, for example, $w = \frac{1}{2}(1 + \sqrt{1 + 4k})$. In particular, if $k = -1$, $w = \frac{1}{2}(\sqrt{3}i + 1)$, and the set of $a + bw$, $a, b \in \mathbb{Z}$, represents a “grid” in the complex plane, see Fig. 4.7.

In fact, one can prove that this is all: *every* quadratic ring is isomorphic to one of the above. Such theorems in mathematics are called *classification* theorems, and are extremely useful: for example, if you want to prove that all quadratic rings satisfy some property, you can just verify this property for the examples above case by case, and, by the classification result, you are done!

Classification of Cubic Rings

Can we classify rings of higher order, say, of rank 3, which are also called *cubic rings*? Well, elements of such ring R can be written in the form $x = x_0 + x_1w_1 + x_2w_2$ for some $w_1, w_2 \in R$ and $x_0, x_1, x_2 \in \mathbb{Z}$. Then multiplication $xy = (x_0 + x_1w_1 + x_2w_2)(y_0 + y_1w_1 + y_2w_2)$ would be completely defined if we define w_1w_2, w_1^2 and w_2^2 . Obviously, not every definition works, for example, setting $w_1w_2 = 1$ and $w_1^2 = 0$ would lead to $1 = 1^2 = (w_1w_2)^2 = w_1^2w_2^2 = 0$, a contradiction. “Classifying” all the possible definitions which do not lead to contradiction seems difficult, but it is possible. For this, we need to consider binary cubic forms, that is, expressions of the form $f(x, y) = ax^3 + bx^2y + cxy^2 + dy^3$ with $a, b, c, d \in \mathbb{Z}$. It turns out that, to every such a form f , we can associate the ring $R(f)$ defined as above with rules $w_1w_2 = -ad$, $w_1^2 = -ac + bw_1 - aw_2$, and $w_2^2 = -bd + dw_2 - cw_2$, and these rules never lead to a contradiction. Can different forms f and g define the same (that is, isomorphic) rings? Actually, they can. For any matrix $\gamma = \begin{bmatrix} p & q \\ r & s \end{bmatrix} \in GL_2(\mathbb{Z})$, and any f as above, let $g(x, y) := \frac{1}{\det(\gamma)} f(px + ry, qx + sy)$, where $\det(\gamma) = ps - qr$. Then the rings $R(f)$ and $R(g)$ turn out to be isomorphic. We will say that such f and g are $GL_2(\mathbb{Z})$ -equivalent. Now, there is a theorem [114] stating that if f and g are *not* $GL_2(\mathbb{Z})$ -equivalent, then the corresponding cubic rings $R(f)$ and $R(g)$ are different (not isomorphic), and, conversely, *every* cubic ring is isomorphic to $R(f)$ for some f . This provides the desired classification result for cubic rings.

Classification of Quartic Rings

What about the classification of quartic rings, that is, rings of rank 4? Before 2004, there was no hope to achieve this, but then a fantastic breakthrough of Bhargava [50] made even this possible! Similar to the above, the elements of a quartic ring Z can be written in the form $x = x_0 + x_1w_1 + x_2w_2 + x_3w_3$ for some $w_1, w_2, w_3 \in R$ and $x_0, x_1, x_2, x_3 \in Z$, and we need to specify all products in the form w_iw_j to define the multiplication. If all $w_iw_j = 0$, the ring is called trivial, and it is called non-trivial otherwise. Bhargava showed how to associate every non-trivial quartic ring to a *pair* (A, B) of integral ternary quadratic forms, that is, expressions of the form $ax^4 + bx^3y + cx^2y^2 + dxy^3 + ey^4$ with $a, b, c, d, e \in \mathbb{Z}$. Conversely, different pairs (A, B) corresponds to different rings unless they can be transformed to each other using a matrix $\gamma_1 \in GL_3(\mathbb{Z})$, and a matrix $\gamma_2 \in GL_2^{\pm 1}(\mathbb{Q})$, where $GL_2^{\pm 1}(\mathbb{Q})$

is the set of all 2×2 matrices γ_2 with rational coefficients such that $\det(\gamma_2) = \pm 1$. This implies that all quartic rings are now fully classified.

Theorem 4.7 *There is a canonical bijection between isomorphism classes of non-trivial quartic rings and $GL_3(\mathbb{Z}) \times GL_2^{\pm 1}(\mathbb{Q})$ -equivalence classes of pairs (A, B) of integral ternary quadratic forms where A and B are linearly independent over \mathbb{Q} .*

Inspired by Theorem 4.7, Bhargava [52] later also developed a classification of quintic rings, that is, rings of rank 5, which is an even more impressive result.

Reference

M. Bhargava, Higher composition laws III: The parametrization of quartic rings, *Annals of Mathematics* **159**-3, (2004), 1329–1360.

4.8 The Time it Takes for a Random Walk to Cover the Plane

Random Walks on the Line and on the Plane

One of the fundamental models in mathematics is the random walk. The simplest one-dimensional random walk is a sequence X_0, X_1, X_2, \dots of random integers defined by the rules $X_0 = 0$ and X_{m+1} is equal to either $X_m - 1$ or $X_m + 1$ with equal chances. Such a random walk serves as the simplest mathematical model for many real-life phenomena, for example, it can model the movement of a gas particle in a thin tube, or the price movement of some financial instrument on the market, or X_m can be the capital of a gambler playing in a casino, assuming that each bet is 1 pound and he has a 0.5 winning probability. In the last example, we assume that the gambler can borrow money, which corresponds to $X_m < 0$.

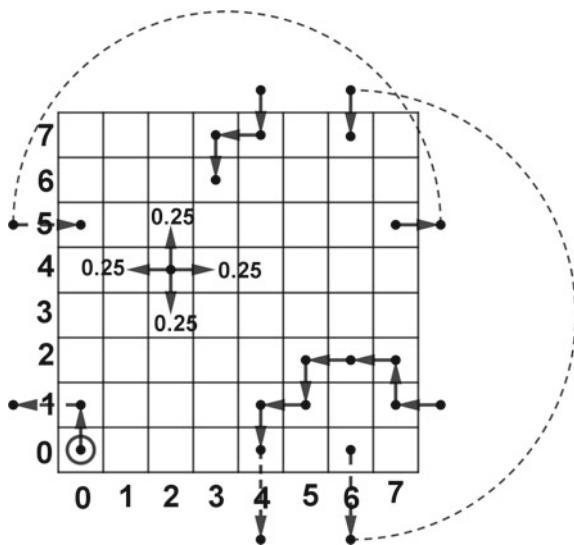
However, we may want to model the joint behaviour of n gamblers, or study the movement of a particle not in a thin tube, but in a three-dimensional space, and in these cases it is natural to extend the definition of random walk to higher dimensions. For $n = 2$, the walk starts at the point $(0, 0)$ in the plane, and then at each step moves one unit up, down, left, or right, with equal probability 0.25. A typical trajectory may look like

$$(0, 0) \rightarrow (0, 1) \rightarrow (-1, 1) \rightarrow (-1, 0) \rightarrow (-2, 0) \rightarrow (-3, 0) \rightarrow (-3, -1) \rightarrow \dots$$

How Long Do We have to Wait Until We Visit Every Point?

A classical theorem states that, if we wait long enough, then, with probability 1, a 2-dimensional random walk will eventually visit *all* points (a, b) with integer coefficients. A fundamental question, however, is *how long* we should wait. For

Fig. 4.8 A random walk on \mathbb{Z}_8^2



example, let T_n be the time needed for a random walk to visit every point (a, b) at distance at most n from the origin, that is, such that $\sqrt{a^2 + b^2} \leq n$. Of course, because the whole process is random, T_n is also a random variable, it may be higher or lower depending on chance, so the real question is to estimate the *distribution* of this random variable, that is, the probability that $T_n \leq k$ for every integer k . For example, what is the probability that $T_1 \leq 10$, that is, that we visit all points $(-1, 0)$, $(0, 1)$, $(1, 0)$, $(0, -1)$ in 10 steps or less? Can we at least be sure that 100 steps suffices with high probability? This is not so obvious. After all, the plane is infinite, so we could start the walk as $(0, 0) \rightarrow (0, 1) \rightarrow (0, 2) \rightarrow (1, 2)$ and go very far in the “positive” direction until eventually coming back to visit $(-1, 0)$.

A Random Walk with a Finite Number of Positions

It seems easier to start thinking about this problem for a random walk with a *finite* number of possible positions. For example, we can define a random walk on the 8×8 chessboard. The squares of the chessboard have coordinates (i, j) for $0 \leq i, j \leq 7$, and the walk is performed by a special “king”, who can move to an adjacent square, but, unlike a real chess king, cannot do diagonal moves. From square $(2, 4)$, he can move to $(3, 4)$, $(2, 5)$, $(1, 4)$, or $(2, 3)$, with the probability of each move being 0.25. To decide this, the king tosses a special die with 4 outcomes: “right”, “up”, “left”, and “down”. However, what if the current position is $(6, 0)$, and the die happened to indicate the move “down”, to square $(6, -1)$, which does not exist on the board?

Imagine the board is a sheet of paper, which we can bend and glue opposite sides together, to form a tube. Then the square one unit “down” from $(6, 0)$ is the opposite square $(6, 7)$, see Fig. 4.8. Similarly, assume that the king reaches $(7, 5)$ and the die

indicates to go further right, to the non-existing square $(8, 5)$. By gluing opposite sides of our tube, and forming a torus, we find that the square to the “right” of $(7, 5)$ is $(0, 5)$. Now every square has exactly four neighbours, and the walk is completely well-defined. A possible walk trajectory may look like

$$(0, 0) \rightarrow (0, 1) \rightarrow (7, 1) \rightarrow (7, 2) \rightarrow (6, 2) \rightarrow (5, 2) \rightarrow (5, 1) \rightarrow (4, 1) \rightarrow (4, 0) \rightarrow (4, 7) \rightarrow \dots$$

see Fig. 4.8.

The Time to Visit All Positions in a Small Lattice Torus

This construction is called the *lattice torus* \mathbb{Z}_n^2 . More generally, the lattice torus \mathbb{Z}_n^2 consists of n^2 squares (or points) with integer coordinates (i, j) for $0 \leq i, j \leq n - 1$, such that points $(0, j)$ and $(n - 1, j)$ are adjacent for any $0 \leq j \leq n - 1$, and the same is true for pairs $(i, 0)$ and $(i, n - 1)$, $0 \leq i \leq n - 1$. Then for $n \geq 2$ any point has exactly four neighbours, and we can define the random walk as above. Now, can we estimate the time T_n needed for a random walk to visit all n^2 points of \mathbb{Z}_n^2 ?

As a simple exercise, let us study the case $n = 2$, where \mathbb{Z}_n^2 consists of just four points $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$. One can easily see that in this case X_m can be either $(0, 0)$ or $(1, 1)$ with equal chances if m is even, and similarly $(1, 0)$ or $(0, 1)$ with equal chances if m is odd. Now, what is the probability that we will never return to $(0, 0)$ after $2k$ steps? This would mean that $X_2 = X_4 = \dots = X_{2k} = (1, 1)$, the probability of which is $1/2^k$. Similarly, the probability that each of the other three points remain uncovered is $1/2^k$, so that the total probability that there exists an uncovered point seems to be $4/2^k$. This is approximately correct, but, to be precise, there is also a small chance $1/2^{2k}$ that *both* points, say, $(1, 0)$ and $(1, 1)$, are uncovered, and this probability was counted twice and should therefore be subtracted. Because there are in total four such pairs of points, we should subtract $4/2^{2k}$, and obtain $4/2^k - 4/2^{2k}$. Because there is no chance for three or more points to be simultaneously uncovered, no further corrections are necessary, and we have $P[T_2 > 2k] = 4/2^k - 4/2^{2k}$. In particular, $P[T_2 > 2] = 1$, (there is no chance to cover four points in two steps), $P[T_2 > 4] = 3/4$, $P[T_2 > 6] = 7/16$, $P[T_2 > 8] \approx 1/4$. We can conclude that the “typical” value of T_2 is about 6.

The Time to Visit All Positions in a Large Lattice Torus

However, this naïve type of analysis becomes complicated even for $n = 3$, while the really interesting case for most application is when n is large, in fact the limiting behaviour of T_n as $n \rightarrow \infty$. Understanding this limiting behaviour turned out to be a surprisingly difficult question. Mathematicians even resolved similar questions in higher dimensions $d \geq 3$, but the seemingly easy 2-dimensional case remained elusive. The theorem we are talking about in this section fully resolves this question

and proves that, for large n , $T_n \approx \frac{4}{\pi}(n \ln n)^2$, or, equivalently, $\frac{T_n}{(n \ln n)^2} \approx \frac{4}{\pi}$. The exact formulation is below.

Theorem 4.8 *If T_n denotes the time it takes for the simple random walk in \mathbb{Z}_n^2 to cover \mathbb{Z}_n^2 completely, then, for any $\varepsilon > 0$,*

$$\lim_{n \rightarrow \infty} P \left[\left| \frac{T_n}{(n \ln n)^2} - \frac{4}{\pi} \right| > \varepsilon \right] = 0.$$

In other words, for any $\varepsilon > 0$,

$$\left(\frac{4}{\pi} - \varepsilon \right) (n \ln n)^2 \leq T_n \leq \left(\frac{4}{\pi} + \varepsilon \right) (n \ln n)^2$$

with probability tending to 1 as $n \rightarrow \infty$. For the chessboard, this approximation gives $T_8 \approx \frac{4}{\pi}(8 \ln 8)^2 \approx 352$.

Extensions to Infinite Walks on the Full Plane and to the Continuous Case

To prove Theorem 4.8, Dembo, Peres, Rosen, and Zeitouni [116] developed a new methodology, a crucial part of which is a control of the probability that two given points are simultaneously uncovered. This methodology turned out to be applicable to a variety of similar problems, including the original one we started with, about the time T_n needed to cover the disk of radius n for an unrestricted random walk on the full plane \mathbb{Z}^2 . In this case, they obtain the result

$$\lim_{n \rightarrow \infty} P(\ln T_n \leq t (\ln n)^2) = e^{-4/t}.$$

We can approximate T_n by a number x such that $P(T_n \leq x) = P(T_n > x) = 0.5$. From the equation $e^{-4/t} = 0.5$ we find $t^* = 4/\ln 2 \approx 5.8$, then $T_n \approx \exp(t^*(\ln n)^2) = n^{t^* \ln n} \approx n^{5.8 \ln n}$. For $n = 4$, the disc contains much fewer than 64 points, but the estimate is $T_4 \approx 4^{5.8 \ln 4} \approx 69000$. This is much larger than the estimate 352 for a similarly sized chessboard. Of course, the estimates are proved for $n \rightarrow \infty$ and there is no reason to expect them to be very accurate for $n = 8$ or $n = 4$, but the inequality $69000 \gg 352$ corresponds to the intuition that on the infinite board it takes *much* longer to cover a given set of points.

The methodology of the authors also works in the continuous case. In particular, they proved that if a particle is moving within the 2-dimensional torus according to the classical model for particle movement, called Brownian motion, then the time C_ε it needs to come within distance ε of every point of a torus is about $\frac{2}{\pi}(\ln \varepsilon)^2$ for small ε . More formally,

$$\lim_{\varepsilon \rightarrow 0} \frac{C_\varepsilon}{(\ln \varepsilon)^2} = \frac{2}{\pi}.$$

In a subsequent work, Belius and Kistler [38] established an even more accurate approximation: $C_\varepsilon = \frac{1}{\pi} \ln \varepsilon (2 \ln \varepsilon - (1 + o(1)) \ln \ln \varepsilon)$, where $o(1)$ is a function which goes to 0 as ε goes to 0.

Reference

A. Dembo, Y. Peres, J. Rosen, and O. Zeitouni, Cover times for Brownian motion and random walks in two dimensions, *Annals of Mathematics* **160**-2, (2004), 433–464.

4.9 On the Geometry of the Uniform Spanning Forest

Building Roads, Graphs, and Trees

Suppose you need to connect n cities by roads, and would like to build as few roads as possible, or you need to connect n computers in a network, by using the minimal number of cables. The mathematical language for this type of problem is that of graph theory – cities or computers are called *vertices* of the graph, and the connections between two of them are called *edges*. Any two vertices may either be connected by an edge or not. If we connect any two vertices by an edge (that is, any two cities by a separate road), this would require $n(n - 1)/2$ edges, but we can obviously do better. For example, for 3 cities A, B, C , it is sufficient to build roads AB and BC , and then the direct road AC is not needed, because people can travel from A to C via B . In general, a graph is called *connected* if, for any two vertices A and B , there exists a path between them, that is, a sequence of edges $AC_1, C_1C_2, \dots, C_{k-1}C_k, C_kB$ for some vertices C_1, \dots, C_k . (If A and B are directly connected by an edge, this is also the path corresponding to $k = 0$.) For example, the graph with vertices A, B, C and edges AB and BC is connected, while the graph with vertices A, B, C and D and edges AB and CD is not connected, because there is no path from A to D .

In this language our problem is to build a *connected* graph with n vertices and as few edges as possible. How to do it? If a candidate solution has a *cycle*, that is, a set of $k \geq 3$ vertices A_1, A_2, \dots, A_k connected by edges $A_1A_2, A_2A_3, \dots, A_{k-1}A_k, A_kA_1$, then it cannot be optimal, because any of these edges can be removed and the remaining graph is still connected. So, we can assume that there are no cycles. A connected graph with no cycles is called a *tree*.

Random Trees as a “Fair” Way to Build Roads

A possible way to connect vertices by a tree is to enumerate them as A_1, A_2, \dots, A_n and then connect them sequentially, that is, using edges $A_1A_2, A_2A_3, \dots, A_{n-1}A_n$. If we use this method for cities, we need $n - 1$ roads, but people from cities A_1 and A_n may complain that it takes too long to travel between their cities. To resolve this, you may select one vertex, say A_1 , and connect it to all others, that is, the edges are

$A_1A_2, A_1A_3, \dots, A_1A_n$. Then this is again a tree, again we need $n - 1$ edges, and now any two vertices A_i and A_j are connected by a path containing at most two edges. However, people from cities A_2, \dots, A_n may complain anyway, asking why city A_1 is in a special position, being connected to every other A_i directly.

A simple theorem in graph theory states that the examples above are indeed optimal, that is, there is no way to connect n cities by fewer than $n - 1$ roads. In fact *every* tree has exactly $n - 1$ edges, so they all work equally well. Unfortunately, out of all trees there is no completely symmetric one, except for the case $n = 2$ in which we can just connect two cities by a road. For example, three cities A, B, C can be connected by $n - 1 = 2$ roads in three ways: (i) AB and AC ; (ii) AB and BC and (iii) AC and BC . In case (i), city A is in the better position, being connected to two other cities, while others are connected only to A . Similarly, in cases (ii) and (iii), cities B and C are in the better position, respectively. Similarly, for $n > 3$, whatever tree we choose, there will always be cities connected to others by exactly 1 road, while to some other cities at least two roads will be built.

How do we resolve this situation fairly? It seems that the only fair way is to organize a lottery, and select the solution at random. For example, in the case of $n = 3$ cities as above, we can choose every tree (i), (ii), or (iii) with equal chance, and then every city has an equal chance to be in the better position. Similarly, for every n , we can select a tree connecting n vertices at random, each with the same probability. This is completely fair, and the number of roads will be $n - 1$, which is optimal.

Random Subtrees of \mathbb{Z}_n^2

However, in practice not only the amount, but also the length of roads are very important. For example, assume that we have $(2n + 1)^2$ cities, located at points (i, j) of the coordinate plane, where $-n \leq i, j \leq n$. Then a random procedure may require building a direct road between cities $(-n, -n)$ and (n, n) , which is obviously not efficient. It makes much more sense to connect only *adjacent* cities, at distance 1 from each other. The resulting tree would not only have $(2n + 1)^2 - 1$ roads, but also the *total length* of all roads is $(2n + 1)^2 - 1$, the minimal possible. More formally, let \mathbb{Z}_n^2 be a graph whose vertices are as above, and edges indicate “possible roads”, that is pairs of cities at distance 1. Then our problem is to select a random tree T which is a subgraph of \mathbb{Z}_n^2 . We will call T a *subtree* of \mathbb{Z}_n^2 .

If the vertices in our graph are not cities but, say, houses in a large city, or all the computers in the internet, or all the atoms in some material, then n could be very large, and the number of possible trees is exponentially larger than n . In these cases, it is useful to study this problem in the limit, as $n \rightarrow \infty$. In the limit, the graph \mathbb{Z}_n^2 become the graph \mathbb{Z}^2 whose vertices are *all* points (i, j) of the coordinate plane with integer coefficients, and the edges are, as earlier, pairs of vertices at distance 1. What would an infinite random subtree T connecting all the vertices of \mathbb{Z}^2 look like? For example, what is the probability that the edge $(0, 0) \rightarrow (0, 1)$ of \mathbb{Z}^2 will be included in T ?

The Uniform Spanning Forest

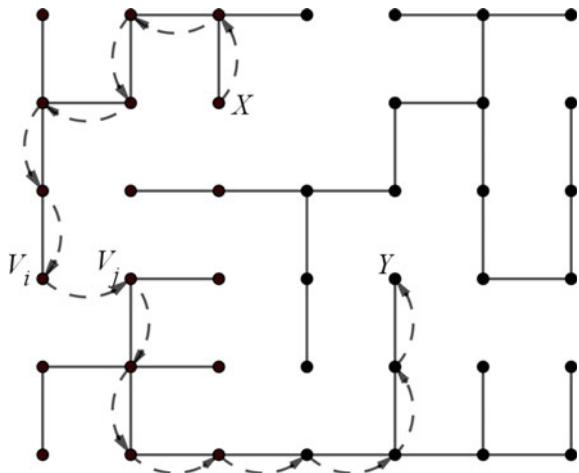
Well, for finite n , the graph \mathbb{Z}_n^2 has $(2n+1)^2$ vertices and $4n(2n+1)$ edges. Indeed, each edge can either be in the form $(i, j) \rightarrow (i, j+1)$ or $(i, j) \rightarrow (i+1, j)$. In the first case, i can be any number from $-n$ to n , and j is any from $-n$ to $n-1$, $(2n+1)2n$ variants in total, and the second case is similar. Out of these edges, we need only $(2n+1)^2 - 1$ for our tree, so that the probability that any given edge is included is about $\frac{(2n+1)^2-1}{4n(2n+1)} = \frac{n+1}{2n+1} = \frac{1+1/n}{2+1/n}$. For large n , we can ignore the small terms $1/n$, and conclude that the probability is about $1/2$. This intuition suggests that any fixed edge, like $(0, 0) \rightarrow (0, 1)$, should be included in a random subtree T of \mathbb{Z}^2 with probability $1/2$.

Then, what does T look like globally? Should we just include any edge of \mathbb{Z}^2 in T with probability $1/2$, independently from each other? Not really. In this case, with probability $1/16$, we would, for example, form a cycle $(0, 0) \rightarrow (0, 1) \rightarrow (1, 1) \rightarrow (1, 0)$, but the probability of such a cycle in any finite tree is 0, not $1/16$. More generally, for any finite set C of edges of \mathbb{Z}^2 , the probability $P(C)$ that all these edges will be included in T should be equal to $\lim_{n \rightarrow \infty} P_n(C)$, where $P_n(C)$ is the corresponding probability in the finite case. It is not at all obvious that it is possible to satisfy these conditions, but in 1991 Pemantle [300] showed that this is indeed the case, and denoted the resulting random graph USF , standing for “uniform spanning forest”. Why “forest”? Because the definition above implies that the probability that it contains a cycle is 0, but, in general, it is not obvious that USF is connected. A graph with no cycles can consist of several separate trees, and is called a *forest*. Figure 4.9 depicts a forest which consists of three separate trees. For example, there is no path from vertex X to vertex Y , which consists only of edges of the graph. In particular, in the “candidate” path depicted in the figure, edge $V_j V_i$ is missing. In general, for any vertices X and Y , let $N(x, y)$ be the minimal number of “missing edges” over all possible paths from X to Y . In Fig. 4.9, $N(X, Y) = 1$.

The Uniform Spanning Forest in Higher Dimensions

In fact, Pemantle proved that, in the 2-dimensional case described above, USF happened to be a tree with probability 1, hence the USF model can indeed be used to “generate” random trees in the plane, and, in particular, build roads between cities in a “fair” way. However, for other applications, it is important to have a similar model in any other dimension d . Unfortunately, the same construction in any dimension $d > 4$ results in a USF which, with probability 1, consists of an infinite number of separate trees. That is, there exist vertices X and Y such that *any* path from X to Y contains some edges outside of USF . In other words, $N(X, Y) > 0$ for some X and Y . Given this, an important research question was how large $N(X, Y)$ can be, that is, how “far” trees in USF are located from each other. The following theorem of Benjamini, Kesten, Peres, and Schramm [39] answers this question in every dimension d .

Fig. 4.9 A fragment of a spanning forest



Theorem 4.9 With probability 1, $N(X, Y) \leq \frac{d-1}{4}$, and there exist pairs X, Y for which $N(X, Y)$ is the largest integer not exceeding $\frac{d-1}{4}$.

In particular, Theorem 4.9 implies that, in dimensions $5 \leq d \leq 8$, USF consists of infinitely many trees T_1, T_2, \dots , but for each T_i, T_j there are vertices $V_i \in T_i$ and $V_j \in T_j$ at distance 1 from each other. Thus, for any $X \in T_i$ and $Y \in T_j$, we can travel from X to V_i in T_i , then make just one “illegal” move from V_i to V_j , and then move from V_j to Y in T_j , as in Fig. 4.9. Starting from dimension 9, we will be forced to make at least 2 “illegal” moves, and so on.

Reference

- I. Benjamini, H. Kesten, Y. Peres, and O. Schramm, Geometry of the uniform spanning forest: Transitions in dimensions 4, 8, 12, ..., *Annals of Mathematics* **160**-2, (2004), 465–491.

4.10 A Polynomial Time Algorithm for Primality Testing

Listing All Small Primes and the Sieve of Eratosthenes

Given a natural number n , can you quickly determine whether there is a non-trivial factorization of n , that is, natural numbers a and b greater than 1 such that $n = ab$? If such a factorization exists, n is called *composite*, otherwise it is called *prime*. For small n , it is easy to determine whether it is composite or prime. For example, 21 is composite because $21 = 3 \cdot 7$, while 37 is prime.

A method to list all prime numbers from 2 to some N has been known since at least the 3rd century BC, and is called the Sieve of Eratosthenes. First, we should

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Fig. 4.10 The Sieve of Eratosthenes for $N = 100$

list all numbers from 2 to N in a table, mark the first one (number 2) as a prime, and then cross out all the numbers which are divisible by 2. Then we claim that the first uncrossed number (which is 3) is prime, cross out all the numbers divisible by 3, and so on, see Fig. 4.10. In fact, after crossing out multiples of just 2, 3, 5, and 7, we have determined all primes from 1 to 100. In general, to list all primes from 1 to N it suffices to cross out only the multiples of primes not exceeding \sqrt{N} . This is because if $n \leq N$ and $n = ab$ is composite, then either $a \leq \sqrt{N}$ or $b \leq \sqrt{N}$.

A Direct Method for Checking if a Number is Prime or Composite

Similarly, to check that a given number n is prime, it suffices to check that it has no divisors between 2 and \sqrt{n} . For, say, $n = 37$, this reduces to checking that $37/2$, $37/3$, $37/4$, $37/5$, and $37/6$ are not integers, which is a straightforward exercise. For $n = 193651$, a by-hand verification of whether any n/i is an integer for $2 \leq i \leq \sqrt{193651} \approx 440$ would require several hours of work (and the result would be that $193651 = 197 \cdot 983$, a composite), but a computer could do this within a second.

However, what if n is a 100-digit number? Then we need to try all possible divisors up to $\sqrt{n} \approx 10^{50}$. Even if a computer could try 10^{10} candidates within a second, it still would take it about 10^{40} seconds, which is more than $3 \cdot 10^{32}$ years. In general, for a k -digit integer the number of operations required by this naïve algorithm is about $\sqrt{10^k} = 10^{k/2}$.

Some Indirect Methods

Fortunately, the primes have a number of useful properties which allow us to test primality indirectly, without finding an explicit factorization. For example, for any integer x , if $x^2 - 1 = (x - 1)(x + 1)$ is a multiple of prime p , then either $x - 1$ or $x + 1$ should be a multiple of p . Hence, if we find at least one x such that this property fails, then p is composite. For instance, $984^2 - 1$ is a multiple of 193651, while 983 and 985 are not, therefore $x = 984$ “certifies” that 193651 is composite. However, it is not clear how to find such a “certification” fast.

Another useful property of primes is Fermat’s Little Theorem, which states that if p is a prime then $a^{p-1} - 1$ is a multiple of p for all natural numbers $1 \leq a \leq p - 1$. For example, $1^4 - 1 = 0$, $2^4 - 1 = 15$, $3^4 - 1 = 80$, $4^4 - 1 = 255$ are all multiples of 5. By Fermat’s Little Theorem, if we can find $a < n$ such that $a^{n-1} - 1$ is not a multiple of n , then n is guaranteed to be composite. For example, $3^9 - 1 = 19682$ is not a multiple of 10, hence 10 is composite.

The Method of Repeated Squares

At first glance, this method is even more complicated than direct division. For example, it is easy to show that 100 is composite because $100 = 10 \cdot 10$, but how to check whether, say, $3^{99} - 1$ is a multiple of 100? After all, 3^{99} is huge... However, to check whether a number is a multiple of 100, it is not necessary to calculate the full number, it is enough to keep track of the remainder after division by 100 (that is, the last 2 digits). Denoting “the remainder after division by 100” as \rightarrow , we can check that $3 \rightarrow 3$, $3^2 \rightarrow 9$, $3^4 \rightarrow 9^2 = 81$, $3^8 \rightarrow 81^2 \rightarrow 61$, $3^{16} = (3^8)^2 \rightarrow 61^2 \rightarrow 21$, $3^{32} = (3^{16})^2 \rightarrow 21^2 \rightarrow 41$, $3^{64} = (3^{32})^2 \rightarrow 41^2 \rightarrow 81$, $3^{96} = 3^{64} \cdot 3^{32} \rightarrow 81 \cdot 41 \rightarrow 21$, $3^{99} \rightarrow 21 \cdot 3^3 \rightarrow 67$, hence $3^{99} - 1$ is not a multiple of 100, which, by Fermat’s Little Theorem, certifies that 100 is composite. This calculation is known as the “method of repeated squares”.

For 100, this method is still more complicated than direct factorization, but it becomes much more useful for larger n . For example, if n is a huge number such as $n = 2^{1000} + 1$, and $a = 3$, then, by an argument similar to the above we can calculate the remainder after division by n for $3, 3^2, 3^4, 3^8$, and so on, arriving at $3^{2^{1000}} = 3^{n-1}$ just after 1000 multiplications! If $3^{n-1} - 1$ is not a multiple of n , we can conclude that n is composite. This is much-much-much quicker than trying about $\sqrt{n} \approx 2^{500}$ possible divisors!

Looking for “Witnesses” of Being Composite

However, what if the calculation shows that $a^{n-1} - 1$ is a multiple of n ? Then Fermat’s Little Theorem cannot tell whether n is composite or prime. However, if $n - 1$ is even, then $a^{n-1} - 1 = x^2 - 1$ for $x = a^{(n-1)/2}$. Then we can check if $x - 1$ or $x + 1$ are multiples of n : if neither are, then n is composite. If $x - 1$ is a multiple

of n and $(n - 1)/2$ is even, then $x - 1 = y^2 - 1$ for $y = a^{(n-1)/4}$, and we can check whether $y - 1$ or $y + 1$ are multiples of n . This process may then be iterated. If it succeeds at some point, we conclude that n is composite and will call such a a *witness* of this fact. If not, we can just try another a .

Miller and Rabin [276, 311] proved that if n is composite then, by this procedure, at least half of all numbers $1 \leq a \leq n - 1$ are witnesses. However, what if n is a 100-digit composite number, and the first witness is, say, a 40-digit number? If we try $a = 2, 3, \dots$ in the increasing order, it would take us ages to find a witness! Conversely, if a 100-digit integer n is prime, then no witness exists, but we cannot be sure about this until we try more than half of all possible candidates.

Using Randomness for Witness Search

At this point, Miller and Rabin arrived at the crucial idea: why not select a *at random* between 1 and $n - 1$? If n is composite and at least half of all numbers in this range are witnesses, then we would find a witness in the first attempt with probability at least $1/2$, and fail with probability at most $1/2$. If we fail, we could try once again, and the probability that we will fail twice is at most $(1/2)^2 = 1/4$. After 100 attempts, the probability that we will fail every time is at most $(1/2)^{100}$, which is essentially impossible. Hence, if we fail 100 times, we can safely conclude that no witnesses exist and n is prime. This procedure is called the Miller–Rabin primality test. It either finds a witness that the number is composite, or concludes that it is prime. The “composite” answer is always correct, and the “prime” answer may be wrong, but with some extremely small probability, and this probability of error can be made as small as we want.

What is important is that the Miller–Rabin test is fast: if n is a k -digit number, and we want the probability of error to be less than 2^{-m} , the number of operations of the algorithm is about mk^2 , which for large k is *much* faster than the $10^{k/2}$ operations for the naïve method. Using the Miller–Rabin test or similar methods, computer packages like Mathematica can easily test 100-digit or even 1000-digit numbers for primality.

A Primality Test Without Randomness

While the Miller–Rabin test solves the problem practically, it does not address the big theoretical question: does there exist a *deterministic* efficient primality test? Here, “deterministic” means an algorithm which does not contain random steps at all, and “efficient” means that any k -digit number should be tested for primality after no more than $Q(k)$ steps, where Q is some fixed polynomial. The class of problems for which there exists a deterministic algorithm working in time bounded by a polynomial from the size of input, is called P . The problem of determining whether a given integer is prime or composite is called PRIMES. With this terminology, the big open question was “does the problem PRIMES belong to the class P ?”.

In 2004, this question was answered by Agrawal, Kayal, and Saxena [5].

Theorem 4.10 *There exists a deterministic polynomial-time algorithm which determines whether a given integer is prime or composite. In other words, PRIMES is in P.*

In the original paper [5], the running time for the algorithm was about $k^{21/2}$ for a k -digit number. Then this bound was improved by other authors, but the algorithm is still much slower than the Miller–Rabin test, and therefore has little practical importance. However, this result is still a big theoretical breakthrough. In fact, there is a big question which asks whether *every* problem which has a randomized polynomial-time algorithm (the class of such problems is called BPP) also has a deterministic one, that is, is $P = BPP$? For many years, PRIMES was the main example of a problem in BPP which was not known to be in P . Theorem 4.10 therefore provides evidence towards the $P = BPP$ conjecture.

The proof of Theorem 4.10, although too long to be presented here, is rather elementary by modern mathematical standards, and uses simple and intuitive ideas. Moreover, two out of the three authors were undergraduate students at the time they proved this result!

Reference

M. Agrawal, N. Kayal, and N. Saxena, PRIMES is in P , *Annals of Mathematics* **160**-2, (2004), 781–793.

Chapter 5

Theorems of 2005



5.1 Structured Additive Patterns in Sets of Positive Density

Comparing Infinite Sets

Can you say that one infinite subset of the set \mathbb{N} of positive integers is larger, or more dense than another one? For example, are there “more” even numbers than multiples of 3? A standard approach in set theory claims that two sets have equal cardinality if there is a one-to-one correspondence between their elements. To every even number $2k$, $k \in \mathbb{N}$, we can put into correspondence a multiple of three $3k$, and vice versa ($2 \leftrightarrow 3$, $4 \leftrightarrow 6$, $6 \leftrightarrow 9$, \dots), hence a set-theorist would say that these two sets are “equally large”. Still, intuitively, even numbers look “denser” in \mathbb{N} than multiples of three, because every second integer is even, while only every third integer is divisible by three. However, how do we formalize this intuition that one infinite set is “bigger” than another one by a factor of $\frac{3}{2}$?

Well, one can argue that \mathbb{N} is the disjoint union of odd and even numbers, odd numbers are “obviously” as dense as even, hence even numbers are “half” of \mathbb{N} . This intuition can be formalized as follows. For $A \subset \mathbb{N}$ and integer n we denote by $A + n$ and call the *translate* of A the set of all numbers representable in the form $a + n$, $a \in A$. For example, if A is the set of even numbers then the set of odd numbers is a translate of A , namely $A - 1$. Hence, the set \mathbb{N} of all natural numbers can be fully covered by two translates of A : $A + 0$ and $A - 1$. On the other hand, if B is the set of multiples of three then we need three translates of B to cover \mathbb{N} , namely B , $B - 1$ and $B - 2$. Hence, A is “larger” than B by a factor of $\frac{3}{2}$.

Syndetic Sets

We denote by $t(A)$ the minimum number of translates of A needed to cover the whole set \mathbb{N} of positive integers, with the convention that $t(A) = \infty$ if no finite number of translates is sufficient. For example, if A is a finite set then obviously $t(A) = \infty$. A set

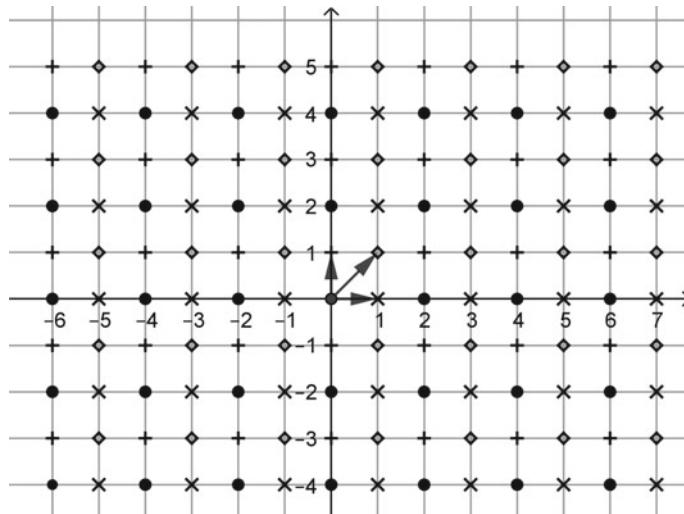


Fig. 5.1 The set A of all points with even coordinates is syndetic with $t(A) = 4$

A for which $t(A) < \infty$ is called *syndetic*. The definition of $t(A)$ and syndetic sets can be straightforwardly extended to subsets of (not necessarily positive) integers \mathbb{Z} , and also to arbitrary dimensions. Let \mathbb{Z}^k be the set of all vectors $x = (x_1, x_2, \dots, x_k)$ with integer coordinates. For any subset $A \subset \mathbb{Z}^k$ and $x \in \mathbb{Z}^k$ we can define the translate $A + x$ to be all the vectors in the form $a + x = (a_1 + x_1, \dots, a_k + x_k)$, $a \in A$, $t(A)$ to be the minimal number of translates of A needed to fully cover \mathbb{Z}^k , and syndetic sets to be the sets with $t(A) < \infty$. For example, if A is the subset of \mathbb{Z}^2 consisting of vectors with both coordinates being even, then A is syndetic with $t(A) = 4$, because \mathbb{Z}^2 is fully covered by the sets A , $A + (0, 1)$, $A + (1, 0)$ and $A + (1, 1)$, see Fig. 5.1. This corresponds to the intuition that A is one fourth of \mathbb{Z}^2 .

However, this definition is appropriate only for “sufficiently regular” sets, even in one dimension. For example, let A be the set of integers whose decimal expansion does *not* start with 9 (that is, the first digit is between 1 and 8). Intuitively, A contains about $8/9$ of all integers, that is, it is much “denser” than the set of even numbers. However, unlike the set of even numbers, the set A is quite “irregular” and has large “gaps”: it contains no integer in the interval $[90, 99]$ of length 10, no integer in the interval $[900, 999]$ of length 100, and, more generally, no integer in the interval $[9 \cdot 10^k, 10^{k+1} - 1]$ of length 10^k . Using this, one can prove that the set \mathbb{N} cannot be covered by any finite number of translates of A , that is, A is not even a syndetic set.

Density and Upper Density

A more general approach to measuring the “density” of a subset $A \subset \mathbb{N}$ is the following. Let $\delta_N(A)$ be the proportion of elements of A among the first N positive

integers (that is, the number of integers $a \in A$ such that $1 \leq a \leq N$, divided by N). If there is a limit $\delta(A) = \lim_{N \rightarrow \infty} \delta_N(A)$, then $\delta(A)$ is called the *density* of A in \mathbb{N} . With this definition, the density of the set of even (and odd) numbers is $1/2$, while the density of the set of multiples of 3 is $1/3$, as expected.

The problem is that this limit is not guaranteed to exist. For example, if A , as above, is the set of integers whose decimal expansion does not start with 9, then $\delta_N(A) = 8/9$ for $N = 10^k - 1$, $k \in \mathbb{N}$, but $\delta_N(A) \approx 80/81$ for $N = 9 \cdot 10^k - 1$, $k \in \mathbb{N}$ (check!). In other words, the sequence $\delta_N(A)$ may have subsequences converging to different limits. In this case, out of these limits of subsequences we may select the *maximal* one, which is denoted \limsup , and then the number $\bar{\delta}(A) := \limsup_{N \rightarrow \infty} \delta_N(A)$ is called the *upper density* of the set A . The advantage of the notion of upper density is that it always exists, in particular, $\bar{\delta}(A) = 80/81$ in the example above.

Because $\bar{\delta}(A + x) = \bar{\delta}(A)$ for every $A \subset \mathbb{N}$ and $x \in \mathbb{Z}$, and $\bar{\delta}(A \cap B) \leq \bar{\delta}(A) + \bar{\delta}(B)$ for every $A, B \subset \mathbb{N}$, $1 = \bar{\delta}(\mathbb{N}) \leq t(A)\bar{\delta}(A)$ whenever $t(A) < \infty$, hence $\bar{\delta}(A) \geq 1/t(A)$. In particular, every syndetic set has a positive upper density. As we have seen above, the converse is not true: a set may have a positive upper density without being syndetic.

Structured Subsets of Sets of Positive Upper Density

An important line of mathematical research is what kind of “structure” is guaranteed to exist in any set of positive upper density. For example, one of the greatest achievement of 20th century mathematics is Szemerédi’s theorem [365] that any set A with $\bar{\delta}(A) > 0$ contains arbitrarily long arithmetic progressions, that is, for any $k \in \mathbb{N}$ there exist integers $a, n \in \mathbb{N}$ such that all $a, a + n, \dots, a + (k - 1)n$ belong to A .

A generalized arithmetic progression of dimension 2 is a set of integers representable in the form $a + e_1 n_1 + e_2 n_2$, where a, n_1, n_2 are fixed integers, and e_1, e_2 take all possible integer values between 0 and some K . The simplest example $K = 1$ results in a quadruple

$$a, a + n_1, a + n_2, a + n_1 + n_2. \quad (5.1)$$

In other words, (5.1) defines quadruples of integers such that the sum of the first and the last integer is the same as the sum of two middle ones ($a + (a + n_1 + n_2) = (a + n_1) + (a + n_2)$). Specific examples of such quadruples are $(1, 2, 3, 4)$, $(1, 2, 4, 5)$, $(1, 2, 5, 6)$, $(1, 2, 6, 7)$, $(1, 3, 4, 6)$, $(1, 3, 6, 8)$, $(2, 5, 10, 13)$, and so on. Would you expect any set A of positive upper density $\delta > 0$ to contain such a quadruple?

An Informal Argument

Here is an informal argument to support this. Let us fix n_1, n_2 and select an integer a at random between 1 and N for some large N with $\delta_N(A) \approx \delta$. Then $a \in A$ with prob-

ability about δ , and similarly $a + n_1 \in A$, $a + n_2 \in A$, and $a + n_1 + n_2 \in A$ with about the same probability. If these four events were independent, then A would contain a full quadruple (5.1) with probability about δ^4 . In general, if something happens with probability 0.001, we would expect it to happen once in 1000 experiments. Similarly, if something happens with probability δ^4 , we would expect it to happen once in $1/\delta^4$ experiments. Hence, if we select N much larger than $1/\delta^4$, and repeat the above experiment about $1/\delta^4$ times, we have a good chance to find a quadruple (5.1) in A .

Unfortunately, this informal argument is not a proof, because events $a \in A$ and $a + n_1 \in A$ are very far from being independent. For example, if A is the set of even integers and n_1 is odd, then the fact that $a \in A$ guarantees that $a + n_1 \notin A$, hence no quadruple (5.1) with odd n_1 (and similarly with odd n_2) can exist in A .

The k -dimensional generalization of (5.1) with parameters n_1, \dots, n_k consists of 2^k integers in the form

$$a + \sum_{i=1}^k e_i n_i, \quad (5.2)$$

where each of e_1, e_2, \dots, e_k is either 0 or 1. An informal argument as above states that a set A with upper density $\bar{\delta}(A) > \delta > 0$ “should” contain infinitely many such progression, in fact the density of such progressions in A is expected to be at least δ^{2^k} . However, as we have seen in the above example with $k = 2$, this intuition is wrong for some parameters n_1, \dots, n_k .

A Formal Theorem

Let us call a set of parameters (n_1, \dots, n_k) *good* if this intuition is *not* wrong, that is, if we can find an integer a such that all 2^k elements of the k -dimensional progression (5.2) belong to A , and, moreover, the set of such a has upper density at least δ^{2^k} . In fact, it is very far from obvious that A contains even *one* such progression, and even less obvious that there exists at least one good set of parameters. The following theorem of Host and Kra [206] proves much more.

Theorem 5.1 *For any $A \subset \mathbb{Z}$ with positive upper density, and any integer $k \geq 1$, the set of good parameters (n_1, \dots, n_k) is syndetic in \mathbb{Z}^k .*

Once again: all we know about the set A is that it has a positive upper density. This density may be very small, and the set itself may be highly “irregular”. However, whatever it is, and no matter how big k is, A is always guaranteed to contain 2^k elements which form a highly symmetric structure (5.2). Moreover, there are “good” parameters (n_1, \dots, n_k) for which it contains this structure for infinitely many values of a , in fact as many as we would expect from the naïve argument. Finally, there are in fact “many” choices of good parameters, so many that a finite number of translations of this set covers the whole of \mathbb{Z}^k .

Measure-Preserving Transformations and Convergence of “Nonconventional Averages”

Theorem 5.1 is a corollary of deep theorem about “measure-preserving transformations”. Let X be, for example, the interval $[0, 1]$, and let T be a function from X to itself, for example, $T(x) = x$ or $T(x) = x^2$. T is called *invertible* if $T(x) = T(y)$ implies $x = y$. For any set $A \subset [0, 1]$, denote by $T^{-1}(A)$ the set of all $x \in [0, 1]$ such that $T(x) \in A$. For example, for $T(x) = x^2$ and $A = [0.25, 0.64]$, we have $T^{-1}(A) = [\sqrt{0.25}, \sqrt{0.64}] = [0.5, 0.8]$. We call a set $A \subset [0, 1]$ measurable if its length (which is also called its “Lebesgue measure” and denoted μ) is well-defined. For example, $\mu(A) = 0.64 - 0.25 = 0.39$ for $A = [0.25, 0.64]$. Next, T is called a *measure-preserving transformation* if $\mu(T^{-1}(A)) = \mu(A)$ for all measurable A . For example, $T(x) = x^2$ is not measure-preserving, because, for $A = [0.25, 0.64]$, $\mu(T^{-1}(A)) = \mu([0.5, 0.8]) = 0.3$ while $\mu(A) = 0.39 \neq 0.3$. On the other hand, $T(x) = x$ is obviously measure-preserving. Another example of a measure-preserving transformation is $T(x) = x + \sqrt{2} - [x + \sqrt{2}]$, where $[y]$ denotes the largest integer not exceeding y . For example, $T(2.2 - \sqrt{2}) = 2.2 - [2.2] = 2.2 - 2 = 0.2$.

Let $L^2(X)$ denote the set of all functions $g : X \rightarrow \mathbb{R}$ such that $\int_X |g(x)|^2 \mu(dx)$ exists and finite. This non-negative real number is called the *norm* of g and denoted by $\|g\|$. We say that a sequence of functions g_1, g_2, \dots converges in $L^2(X)$ to a function $g \in L^2(X)$, and write $g = \lim_{N \rightarrow \infty} g_n$, if $\lim_{N \rightarrow \infty} \|g - g_N\| = 0$.

For any measure-preserving transformation $T : X \rightarrow X$, and any $x \in X$, denote by $T^2(x)$ the number $T(T(x))$, $T^3(x)$ the number $T(T(T(x)))$, and, more generally, $T^n(x)$ the number $T(T(\dots T(x) \dots))$, where T is iterated n times. For example, if $X = [0, 1]$ and $T(x) = x^2$, then $T^2(x) = (x^2)^2 = x^4$, $T^3(x) = ((x^2)^2)^2 = x^8$, and $T^n(x) = x^{2^n}$. The main theorem of Host and Kra [206] establishes the existence of the following limit in $L^2(X)$

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} f_1(T^n(x)) f_2(T^{2n}(x)) \dots f_k(T^{kn}(x))$$

for any invertible measure-preserving transformation T , any integer k , and any measurable functions f_1, f_2, \dots, f_k . Previously, it was known that this limit exists and is constant (that is, the same for all x) for some “special” transformations T . In general, however, the limit may depend on x , (in this case, the expressions under the limit are called “nonconventional averages”), and proving convergence in this case becomes much more difficult. The authors did this by using properties of some geometric objects called “nilmanifolds”. Theorem 5.1 is one of many corollaries of this deep result.

Reference

B. Host and B. Kra, Nonconventional ergodic averages and nilmanifolds, *Annals of Mathematics* **161**-1, (2005), 397–488.

5.2 A Sharp Form of Whitney's Extension Theorem

“Restoring” a Function From a Finite Set of Its Values

Assume that you need to measure the radiation level along a street. You have a device which can show you the radiation level at any point you like. Of course, you cannot do direct measurements at *every* point along the street—there are infinitely many of them. What you can do, is measure it at some finite set of points, and then “extend” the result to the full street.

Let x_1, x_2, \dots, x_n be the coordinates of the points at which you performed the measurements, and let y_1, y_2, \dots, y_n be the corresponding levels of radiation. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a function representing the level of radiation, that is, the radiation at the point with coordinate x is $f(x)$. Then you know that $f(x_i) = y_i, i = 1, 2, \dots, n$, and would like to “restore” the function f at all the remaining points based on this information.

What Kind of f Do We Want? Continuous and Bounded

Of course, there are infinitely many ways to do this, but some of them look “nicer” than others. To start, let us consider an extremely simple example, in which you took measurements at points 0, 1, and 2, and obtained results $f(0) = 0$, $f(1) = 1$, and $f(2) = 0$. Then, a possible “extension” would be a function f such that $f(x) = x$ for $x \neq 2$ but $f(2) = 0$, see Fig. 5.2a. Obviously, this “extension” looks unsatisfactory for a number of reasons. First of all, it is natural to expect that f should be a continuous function. Secondly, there is no reason at all to believe that the radiation increases in an unbounded fashion across the street. In fact, because all your measurements returned a value between 0 and 1, why not assume that $0 \leq f(x) \leq 1$ for all x ?

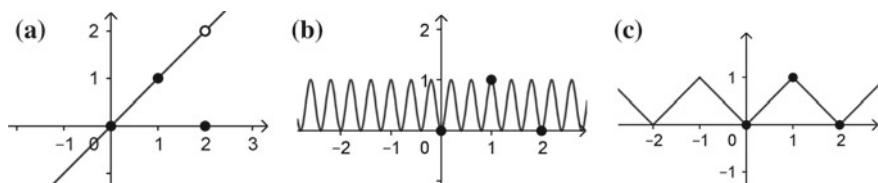


Fig. 5.2 Some functions f such that $f(0) = 0$, $f(1) = 1$, and $f(2) = 0$

What Kind of f Do We Want? Fluctuating Not Too Much

One of the simplest examples of a continuous bounded function $f : \mathbb{R} \rightarrow \mathbb{R}$ is $f(x) = \sin x$, or, more generally, $f(x) = A + B \sin(Cx + D)$ for some constants A, B, C, D . It takes values between $A - B$ and $A + B$, and we want it to take values between 0 and 1, hence we can choose A and B such that $A - B = 0$ and $A + B = 1$, that is, $A = B = 1/2$. Next, to satisfy conditions $f(0) = f(2) = 0$ and $f(1) = 1$ we can select $D = -\pi/2$ and $C = \pi l$ for any odd integer l , resulting in

$$f(x) = \frac{1}{2}(1 + \sin[-\pi/2 + \pi l x]). \quad (5.3)$$

However, this example also looks “strange” for large l : such a function would increase fast on $[0, 1/l]$, then decrease fast on $[1/l, 2/l]$, and so on, see Fig. 5.2b for $l = 5$. There is nothing in our data indicating behaviour like this. The selection $l = 1$ looks much more natural.

Problem Formulation: First Attempt

To summarize, we would like our function f to be continuous, take values not too big and not too small, and also increase and decrease not too fast. The latter condition can be formalized by saying that the *derivative* of f should not be too big or too small. The derivative $f'(x)$ formalizes the notion of the “rate” of increase/decrease of a function f at the point x , and is defined by $f'(x) := \lim_{\varepsilon \rightarrow 0} (f(x + \varepsilon) - f(x))/\varepsilon$, see Sect. 3.1 for a more detailed discussion of this concept. For every differentiable function $f : \mathbb{R} \rightarrow \mathbb{R}$ its C^1 norm is

$$\|f\|_{C^1} := \max\{\sup_{x \in \mathbb{R}} |f(x)|, \sup_{x \in \mathbb{R}} |f'(x)|\}.$$

For example, for f given by (5.3), $f'(x) = \frac{\pi l}{2} \cos[-\pi/2 + \pi l x]$, and $\sup_{x \in \mathbb{R}} |f'(x)| = \frac{\pi |l|}{2}$, which can be very large for large l .

Now we can formulate our question: given data x_1, x_2, \dots, x_n and y_1, y_2, \dots, y_n , and a constant $M > 0$, can we find a function $f : \mathbb{R} \rightarrow \mathbb{R}$ such that (i) $f(x_i) = y_i$, $i = 1, 2, \dots, n$ and (ii) $\|f\|_{C^1} \leq M$? Condition (i) states that the function agrees with our data, while condition (ii) states that it is sufficiently “nice”.

Necessary and Sufficient Conditions for the Existence of a Solution

Of course, if there exists a y_i such that $|y_i| > M$, then the task is impossible: indeed, by definition of the C^1 norm, $\|f\|_{C^1} \geq |f(x_i)| = |y_i| > M$. In words, if some of our data exceed the threshold M , we cannot expect that $\|f\|_{C^1}$ will not exceed it.

Also, if there exist two points x_i and x_j such that $|y_i - y_j| > M|x_i - x_j|$ then $\|f\|_{C^1} \leq M$ is impossible as well. For example, if $x_i < x_j$ and $y_i < y_j$, condition $f(x_j) - f(x_i) = y_j - y_i > M(x_j - x_i)$ implies that the function f increases with a rate greater than M on the interval (x_i, x_j) , hence there exists at least one point $x \in (x_i, x_j)$ with $f'(x) > M$, which implies that $\|f\|_{C^1} > M$.

Hence, simple necessary conditions for the existence of f satisfying (i) and (ii) are (a) $|y_i| \leq M$, for all $i = 1 \dots, n$, and (b) $|y_i - y_j| \leq M|x_i - x_j|$ for all $i, j = 1 \dots, n$. A very special case of the theorem we will discuss below implies that these conditions would also be sufficient if we made them stronger by constant factor. That is, if there exists a constant A such that (a) $|y_i| \leq M/A$, for all $i = 1 \dots, n$, and (b) $|y_i - y_j| \leq M|x_i - x_j|/A$ for all $i, j = 1 \dots, n$, then f satisfying (i) and (ii) is guaranteed to exist. In other words, if our data are bounded and do not “oscillate” too much, then they can be extended to the full real line in a “nice” way, where “nice” is understood in terms of a small C^1 norm.

Problem Formulation: Second Attempt

However, even a small $\|f\|_{C^1}$ norm does not guarantee that the function f is “nice”. For example, our data $f(0) = f(2) = 0$, $f(1) = 1$, can be extended to a function

$$f(x) = x, \quad 0 \leq x \leq 1, \quad \text{and} \quad f(x) = 2 - x, \quad 1 \leq x \leq 2,$$

which can then be extended to the full real line in a periodic way, see Fig. 5.2c. We can also make it “smooth” by modifying it in small neighbourhoods of integer points, and guaranteeing that $\|f\|_{C^1} \leq 1 + \varepsilon$ for some small $\varepsilon > 0$, which is essentially the best possible for our data. However, this function is nevertheless a bit “strange”: why do we think that the radiation increases at a uniform rate on $0 \leq x \leq 1$, and then *suddenly* starts to decrease after “passing” 1? The problem here is that f' changes fast. The function responsible for the rate of increase/decrease of f' is, of course, the derivative of f' , which is denoted $f^{(2)}$ and called the *second derivative* of f . f' would not increase/decrease too quickly if $f^{(2)}$ were bounded by M . Of course, we may also require that $f^{(2)}$ should not change too quickly, which is equivalent to boundedness of the third derivative $f^{(3)}$, and so on. In general, for every m times differentiable function $f : \mathbb{R} \rightarrow \mathbb{R}$ its C^m norm is

$$\|f\|_{C^m} := \max\{\sup_{x \in \mathbb{R}} |f(x)|, \sup_{x \in \mathbb{R}} |f'(x)|, \dots, \sup_{x \in \mathbb{R}} |f^{(m)}(x)|\}.$$

Now our question becomes: For any positive integer n , data x_1, x_2, \dots, x_n and y_1, y_2, \dots, y_n , and constant $M > 0$, can we find a function $f : \mathbb{R} \rightarrow \mathbb{R}$ such that (i) $f(x_i) = y_i$, $i = 1, 2, \dots, n$ and (ii) $\|f\|_{C^m} \leq M$?

The Solution

The following theorem of Fefferman [144] gives the full solution to this difficult question.

Theorem 5.2 *Given an integer $m \geq 1$, there exists an integer k_m and a constant A_m , depending only on m , for which the following holds. Let $f : E \rightarrow \mathbb{R}$ be a function defined on a finite set $E \subset \mathbb{R}$. Let M be a given, positive number. Suppose that, for any k distinct points $x_1, \dots, x_k \in E$, with $k \leq k_m$, there exist polynomials P_1, \dots, P_k of degree $m-1$ satisfying (a) $P_i(x_i) = f(x_i)$ for $i = 1, \dots, k$; (b) $|P_i^{(u)}(x_i)| \leq \frac{M}{A_m}$ for $i = 1, \dots, k$ and $u = 0, \dots, m-1$; and (c) $|P_i^{(u)}(x_i) - P_j^{(u)}(x_i)| \leq \frac{M}{A_m} |x_i - x_j|^{m-u}$ for $i, j = 1, \dots, k$ and $u = 0, \dots, m-1$. Then there exists an m times differentiable function $F : \mathbb{R} \rightarrow \mathbb{R}$ with $\|F\|_{C^m} \leq M$ such that $F(x) = f(x)$ for all $x \in E$.*

The conditions in Theorem 5.2 look complicated, but are directly verifiable, while the conclusion is exactly what we need: extend the function f from a finite set to the whole real line such that it is not too big, and likewise for its derivative, and the derivative of its derivative, and so on, up to the m -th order. In fact, Fefferman proved a more general version of this theorem, guaranteeing the existence of “nice” extensions of f from a finite set to d -dimensional space \mathbb{R}^d . For example, for $d = 2$ it allows you to measure radiation at a finite number of points within a city, and then extend the result in a nice way to the whole city. This general theorem is a “refined” and “sharpened” version of an old and classical result in this direction called Whitney’s extension theorem [402].

Reference

C. Fefferman, A sharp form of Whitney’s extension theorem, *Annals of Mathematics* **161**-1, (2005), 509–577.

5.3 Minimal Surfaces in 3-Space I

How to Save Asphalt When Asphalting Your Yard

Assume that you need to put asphalt on your yard, which has a circular form with boundary length k meters. How much asphalt would you need? This is easy! The radius of the circle is $r = \frac{k}{2\pi}$, hence its area is $S = \pi r^2 = \pi \left(\frac{k}{2\pi}\right)^2 = \frac{k^2}{4\pi}$ square meters.

This answer is correct for any practical purposes, but, just for fun, let us recall that the actual shape of our planet is not a plane but a sphere! So, technically, you would need to put asphalt on the part of the sphere surrounded by a circle of length k , hence you would need to cover more than $\frac{k^2}{4\pi}$ square meters! For “practical” values of k the difference is tiny, but, again just for fun, assume for a moment that $k = 2\pi R$, where

R is the radius of the sphere. In this case you would need to cover half of the whole area of the sphere, that is, $\frac{1}{2}(4\pi R^2) = 2\pi R^2 = \frac{k^2}{2\pi}$, twice more than our previous $\frac{k^2}{4\pi}$ result!

However, we can “cheat” in the following way: let us imagine a plane which crosses the Earth’s surface exactly at the boundary of the yard. Such a plane would go underground behind the yard, so, technically, we could remove all the clay above the plane, make the yard surface exactly flat, and then $\frac{k^2}{4\pi}$ square meters of asphalt would suffice.

Minimal Surfaces

This discussion reveals the important difference between a plane and a sphere. If we assume that the earth is a flat plane, any cheating is impossible: any attempt to add or remove some clay to/from the yard would only increase the area we need to asphalt.

Let us state this property a little more formally. A surface $M \subset \mathbb{R}^3$ is called a *minimal surface* if and only if every point $p \in M$ has a neighbourhood S , with boundary B , such that S has a minimal area out of all surfaces S' with the same boundary B . The discussion above implies that a plane is a minimal surface, while a sphere is not. Minimal surfaces play an important role in mathematics and physics, because real-world materials like soap films are in states of minimum energy when they are covering the least possible amount of area, and therefore they “try to form” minimal surfaces.

As another example, imagine you need to build a roof for a very large stadium. Sometimes stadiums have a complicated structure, with possibly varying height across the stadium, hence a flat roof is impossible. How much material would you need, and what would the most material-economic roof form be in this case? Mathematically, the question is: for any fixed simple (that is, without self-intersection) closed curve $B \subset \mathbb{R}^3$, we need to consider all possible surfaces S with boundary B , and select the one with minimal area. This is a minimal surface with boundary B , and a roof in this form was indeed constructed for the Olympic Stadium in Munich.

Examples of Minimal Surfaces

A plane is an example of a minimal surface which “goes infinitely far”, and does not have a boundary. One may ask, does there exist another minimal surface without boundary, except the plane? It turns out, there does. One of the simplest examples, described by Euler in 1774, is a *helicoid*. This is the surface which can be defined by equations

$$x = s \cos(\alpha t), \quad y = s \sin(\alpha t), \quad z = t, \quad (5.4)$$

where α is constant, and s, t are real parameters, ranging from $-\infty$ to ∞ . For $t = 0$, this is the line with equation $y = z = 0$. When t increases/decreases, the line

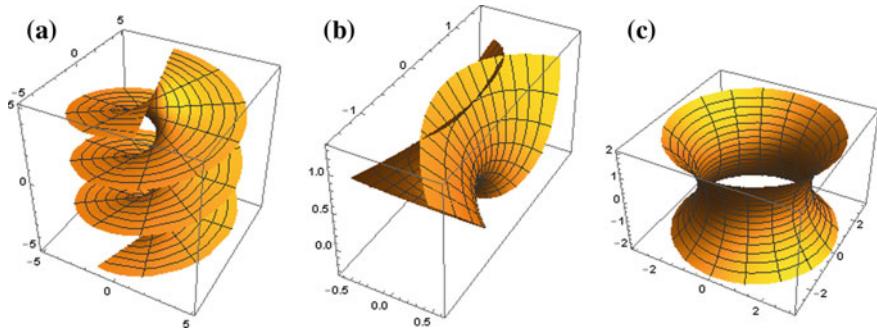


Fig. 5.3 Fragments of **a** a helicoid, **b** the Enneper surface, and **c** a catenoid

is rotated in the (x, y) coordinates, and moves up/down in the z coordinate. This resulting shape looks like a spiral, see Fig. 5.3a.

Since then, various examples of minimal surfaces have been discovered. However, all of them are in some sense not as “nice” as the plane and helicoid. For example, the Enneper surface is a minimal surface given by the equations

$$x = u(1 - u^2/3 + v^2)/3, \quad y = -v(1 - v^2/3 + u^2)/3, \quad z = (u^2 - v^2)/3, \quad (5.5)$$

where u, v are parameters. However, it is possible to find different pairs (u_1, v_1) and (u_2, v_2) leading to the same (x, y, z) in (5.5). Geometrically, this corresponds to self-intersections, see Fig. 5.3b. A surface M is called *properly embedded* in \mathbb{R}^3 if it has no boundary and no self-intersections.

An example of a properly embedded minimal surface in \mathbb{R}^3 is a catenoid, given by equations

$$x = c \cosh \frac{v}{c} \cos u, \quad y = c \cosh \frac{v}{c} \sin u, \quad z = v,$$

where $\cosh(x) := \frac{1}{2}(e^x + e^{-x})$, $c \neq 0$ is a constant, and $u \in [-\pi, \pi]$ and $v \in \mathbb{R}$ are parameters. It was found by Euler in 1744, even before the helicoid. However, while the helicoid looks like a deformed plane, the catenoid looks rather like a deformed cylinder, see Fig. 5.3c.

Simply Connected Surfaces

How can we formally define the difference between a “deformed plane” and a “deformed cylinder”? Imagine you and your friend are two-dimensional “people” living in the plane S . Your friend claims that the surface S you live in is really not a plane but a cylinder. How can you prove to your friend that he is wrong, without leaving the surface?

Here is a solution. You can ask your friend to select any circle he likes, or, more generally, any simple closed curve B . You can then demonstrate that you can contract B to a point in a continuous way without ever leaving S . It is clear that this can always be done within the plane. However, if you were living on a cylinder, your friend could present to you a curve which “goes around” the cylinder, and it would be impossible to contract it to a point in a continuous way without ever leaving the cylinder.

A surface S which has this property (any simple closed curve $B \subset S$ can be contracted to a point) is called *simply connected*. It is easy to see from the pictures that the plane and helicoid are simply connected, while the cylinder and catenoid are not. Intuitively, simply connected surfaces have no “holes” and they are perhaps one of the simplest non-trivial classes of surfaces one may wish to study.

Characterization of the Helicoid

As mentioned above, since 1774, many more examples of minimal surfaces have been discovered. One possible infinite family is, informally, to take a helicoid, and add an arbitrary finite number k of “handles” to it. However, if a surface has a “handle”, it is already not simply connected. For more than 200 years, no-one was able to construct another example of a simply connected properly embedded minimal surface. The following theorem of Meeks and Rosenberg [268] explains why: it turns out that there are no other examples!

Theorem 5.3 *A properly embedded simply connected minimal surface in \mathbb{R}^3 is either a plane or a helicoid.*

Of course, Theorem 5.3 is not the only result distinguishing the plane and helicoid from all other minimal surfaces. For example, a surface S is called *ruled* if through every point of S we can draw a straight line that lies on S . Obviously, the plane is a ruled surface, while (for example) a sphere is not. One can easily check that the helicoid is a ruled surface as well. Indeed, for any point $A_0 = (x_0, y_0, z_0)$ on it, (5.4) implies that the line described by the equations $\sin(\alpha z_0)x - \cos(\alpha z_0)y = 0$ and $z = z_0$ passes through the point A_0 and is fully contained in the helicoid. Catalan proved in 1842 that the helicoid and the plane are the only ruled minimal surfaces.

However, while ruled surfaces are somehow “artificial” (we do not expect a “typical” surface to contain too many straight lines), simply connected surfaces lie at the heart of the field of mathematics called topology, and therefore Theorem 5.3 is the most fundamental characterization of the helicoid one could possibly hope for. It also opens the road to characterizations of minimal surfaces with a more complicated structure.

Reference

W.H. Meeks III and H. Rosenberg, The uniqueness of the helicoid, *Annals of Mathematics* **161**-2, (2005), 727–758.

5.4 Statistical Properties of Quadratic Dynamics

A Non-trivial Question About Quadratic Polynomials

Can you believe that some of the deepest mathematical theorems of the 21st century are about quadratic polynomials of the form

$$f_a(x) = a - x^2,$$

where a is a real parameter? At first glance, you can easily answer any reasonable question about this polynomial, like for what parameters a it has real roots, etc. However, here is a question which is not so easy to answer: take, say, $x_0 = 0$, and form the infinite sequence

$$x_0, \quad x_1 = f_a(x_0), \quad x_2 = f_a(x_1), \dots, x_{n+1} = f_a(x_n), \dots, \quad (5.6)$$

which is an example of a *dynamical system*. Can we describe the behaviour of such a sequence in a satisfactory way?

Cycles, Attracting Cycles, and Regular Quadratic Polynomials

Well, for some parameters a this is easy. For example, for any $a < -1/4$ it is easy to prove¹ that $f_a(x) < x$, $\forall x \in \mathbb{R}$, and any sequence (5.6) just monotonically decreases to $-\infty$. Similarly, for $a > 2$ one can prove that $f_a(x) < x$, $\forall x < -a$. Combining this with the fact that $x_0 = 0$, $x_1 = a$, $x_2 = a - a^2 < -a$, we can conclude that the sequence again decreases to $-\infty$. Hence, we can restrict our attention to the region $a \in [-1/4, 2]$.

Even within this region some parameters are easy to analyse. For example, for $a = 0$ the sequence (5.6) becomes $x_1 = -x_0^2$, $x_2 = -x_1^2 = -x_0^4$, \dots , and the formula for the general term is $x_n = -x_0^{2^n}$. With $x_0 = 0$, this is just a sequence of zeros. More generally, for any $x_0 \in (-1, 1)$, the sequence quickly converges to 0.

For $a = 1$, and $x_0 = 0$, the sequence (5.6) becomes 0, 1, 0, 1, 0, 1, \dots , that is, it returns to 0 after 2 steps, and then repeats itself in a periodic way. If $x_k = x_0$ for some k , the part of the sequence $(x_0, x_1, \dots, x_{k-1})$ is called a *cycle*.

For $x_0 = 0.2$, the first terms of the sequence (5.6) look like

$$0.2, \quad 0.96, \quad 0.0784, \quad 0.9939, \quad 0.0123, \quad 0.9998, \quad 0.0003, \dots$$

that is, it quickly converges to the cycle $(0, 1)$. A cycle $(x_0 = x^*, x_1, \dots, x_{k-1})$ is called *attracting* if there exists an $\varepsilon > 0$ such that the sequence (5.6) converges to this

¹Indeed, $f_a(x) < x$ is equivalent to $a - x^2 < x$, or $x^2 + x > a$, or $x^2 + x + 1/4 > a + 1/4$, or $(x + 1/2)^2 > a + 1/4$, which is obvious for any $a < -1/4$, because the left-hand side is non-negative, while the right-hand side is negative.

cycle for any x_0 such that $|x_0 - x^*| < \varepsilon$. A (quadratic) function f is called *regular* if the corresponding dynamical system (5.6) has an attracting cycle, see Sect. 2.7.

Approximate Periodic Behaviour

As noted in Sect. 2.7, it is easy to analyse the sequence (5.6) for regular f , but the problem is that there exist parameters a for which the function $f_a(x) = a - x^2$ is *not* regular. For example, for $a = 1.5$ the sequence (5.6) starts with

$$0, 1.5, -0.75, 0.9375, 0.6211, 1.1142,$$

$$0.2585, 1.4332, -0.5541, 1.1930, 0.0767, 1.4941 \dots$$

and does not demonstrate any regular behaviour. How can we analyse such a sequence?

Here is an idea which can help at least partially. In the example above, we can notice that $x_{10} = 0.0767$, pretty close to 0. Because $f(x) = 1.5 - x^2$ is a continuous function, the fact that $x_{10} \approx x_0$ implies that $x_{11} = f(x_{10}) \approx f(x_0) = x_1$, then, by a similar argument, $x_{12} \approx x_2$, and so on. Hence, the sequence has a periodic behaviour (at least approximately) at least for some time after every point t such that $x_t \approx 0$.

Searching for x_t Close to 0

In the example above, $x_{11} \approx 1.4941$ is indeed pretty close to $x_1 = 1.5$, but, for example, $x_{16} \approx 0.1230$ differs quite substantially from $x_6 \approx 0.2585$. Intuitively, if we could find a t such that, say, $|x_t| < 0.01$, the periodic behaviour after this point would hold with better accuracy, and stay for longer. And indeed, computation shows that $x_{72} \approx 0.0080 < 0.01$, and then, starting from the 73rd term, the sequence repeats itself quite well, for example, $x_{78} \approx 0.2570$, which is quite close to x_6 , see Fig. 5.4.

Does there exist a term x_t such that, say, $|x_t| < 0.001$, or, more generally, $|x_t| < \varepsilon$ for any $\varepsilon > 0$? And, if so, how many terms of the sequence should we expect to compute before finding it? Intuitively, it should exist, why not? If the sequence takes

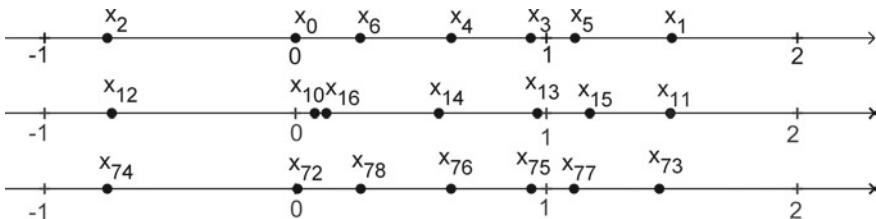


Fig. 5.4 Approximate periodic behaviour of the sequence $x_{n+1} = 1.5 - x_n^2$

values in some unpredictable, “random” fashion on some interval $[b, d]$ containing 0, then the “probability” that any x_t belongs to $[-\varepsilon, \varepsilon]$ is about $2\varepsilon/(d - b)$, hence we can expect to find such a term after computing about $n = (d - b)/2\varepsilon$ of them. To put it differently, after examining n terms we can expect to find one with $|x_n| < \varepsilon$ for $\varepsilon = (d - b)/2n$. To simplify the notation, let us introduce the constant $c = (d - b)/2$, and then $\varepsilon = c/n$, resulting in $|x_n| < c/n$.

Some Difficulties of Rigorous Analysis

There are two problems with turning the intuitive argument above into a formal proof:

- (a) we are talking about a deterministic sequence, so there are no probabilities whatsoever, and
- (b) the calculation above is made using the assumption that the sequence x_n is distributed in an approximately uniform way on the internal $[b, d]$, which is clearly not the case.

The first problem may in principle be addressed using Theorem 2.7 discussed in Sect. 2.7. It states that for almost all parameters $a \in [-1/4, 2]$ the quadratic polynomial $f_a(x) = a - x^2$ is either regular or “stochastic”. Here, by “almost all parameters” we mean “all except a set of measure 0”, that is, a set that can be covered by a collection of intervals of total length ε for any $\varepsilon > 0$; and by “stochastic” we informally mean that we really can think of x_n as a sample from some probability distribution, see Sect. 2.7 for the formal definition.

However, Theorem 2.7 cannot help with problem (b), because x_n can “cover” the interval $[b, d]$ in a highly non-uniform way: there may be, say, many more points near 0.5 than near 0, hence it is far from clear how small a term (in absolute value) of the sequence we can find after examining the first n terms.

The Solution

The following theorem of Avila and Moreira [25] answers this question in a reasonably precise way, answering a long standing conjecture of Sinai.

Theorem 5.4 *For almost all parameters $a \in [-1/4, 2]$ such that $f_a(x) = a - x^2$ is nonregular, and $x_0 = 0$, the set of n such that $|x_n| < 1/n^\gamma$ is finite if $\gamma > 1$ and infinite if $\gamma < 1$.*

Theorem 5.4 implies that our informal intuition described above is essentially correct. For any $\gamma < 1$, for example, $\gamma = 0.99$, we can find an infinite subsequence $x_{n_1}, x_{n_2}, \dots, x_{n_t}, \dots$ such that $|x_{n_t}| < 1/n_t^{0.99}$, $t = 1, 2, \dots$. The fact that $x_{n_t} \approx 0 = x_0$ implies $x_{n_t+1} \approx x_1, x_{n_t+2} \approx x_2$, and so on, and the smaller $|x_{n_t}|$ the better quality of this approximation.

However, if we would like to have an even better “quality” of approximation and guarantee that $|x_{n_t}| < 1/n_t^{1.01}$, $t = 1, 2, \dots$, then we can find only a finite number

of terms in the sequence which just happen to be so unusually small, but no infinite subsequence satisfying this inequality can exist.

Putting it differently, if we would like to find a term such that $|x_t| < \varepsilon$, we really need to investigate about $1/\varepsilon$ terms: if we look at just $1/\varepsilon^{0.99}$ terms, we may be lucky to find such x_t for some values of ε , but in general will not succeed.

The functions $f_a(x) = a - x^2$ studied in Theorem 5.4 have the property that they are increasing for $x < x_0 = 0$, have a maximum at $x = x_0$, and are decreasing after this. Functions having this property (or, conversely, which decrease to a minimum and then increase) are called *unimodal maps*. For example, $f(x) = 1 - x^4$ is a unimodal map, while the function $f(x) = \sin(x)$ on $[-4\pi, 4\pi]$ is not unimodal, because intervals where it increases/decreases alternate several times. Theorem 5.4 is an important step towards proving a similar result for all unimodal maps, not necessary quadratic ones.

Reference

- A. Avila and C. Moreira, Statistical properties of unimodal maps: the quadratic family, *Annals of Mathematics* **161**-2, (2005), 831–881.

5.5 Every Subset of Primes of Positive Density Contains a 3-Term Progression

3-Term Arithmetic Progressions Formed From Some “Special” Integers

A (non-trivial) 3-term arithmetic progression (3AP) is a set of 3 numbers $a < b < c$ such that $b - a = c - b$, or, equivalently, $b = (a + c)/2$. For example, $(1, 2, 3)$, $(2, 4, 6)$ and $(1, 5, 9)$ are arithmetic progressions. The last two examples demonstrate that we can easily form a 3AP with only even, or only odd, integers.

Can we form a 3AP using, for example, only powers of 2? It turns out, we cannot. For three integers $2^k < 2^m < 2^n$ the equation $2^m - 2^k = 2^n - 2^m$ would imply $2^{m-k} - 1 = 2^{n-k} - 2^{m-k}$, which is impossible because the left-hand side is an odd number while the right-hand side is an even one.

Upper Density and the Erdős–Turán Conjecture

This observation is not surprising, because, intuitively, there are “much more” even (or odd) numbers to choose from than powers of 2: while “half” of all integers are even, only a “tiny fraction” of them are powers of 2. And, of course, the more numbers we have to choose from, the “easier” it is to find an arithmetic progression.

We need a bit of care to formally define what we really mean by saying that one infinite set (even numbers) is “larger” than another one (powers of 2), but the correct definition is the following one. For any set A of positive integers, and any positive integer N , we let $|A(N)|$ denote the number of elements of A not exceeding N , and

then the number $\delta(A) := \lim_{N \rightarrow \infty} \frac{|A(N)|}{N}$ is called the *density* of A , provided that the limit exists. It is easy to check that the density of even (and odd) numbers is 0.5, while the density of powers of 2 is 0. For some sets A , the limit in the definition of density may not exist, see Sect. 5.1 for an example and a detailed discussion, but there always exists the number²

$$\bar{\delta}(A) := \limsup_{N \rightarrow \infty} \frac{|A(N)|}{N}, \quad (5.7)$$

which is called the *upper density* of the set A .

Erdős and Turán [138] conjectured in 1936 that any set A with positive upper density (that is, such that $\bar{\delta}(A) > 0$) must contain a 3AP. In other words, any set A containing no 3AP should have $\bar{\delta}(A) = 0$, which implies that the density $\delta(A)$ exists and is equal to 0.

Small Sets Without a 3-Term Arithmetic Progression

Let us first study this question for some small finite sets. How many integers can we select from the set $\{1, 2, \dots, 9\}$ such that the selected subset contains no non-trivial 3-term arithmetic progression? We can surely select 5, for example, $\{1, 2, 6, 7, 9\}$. Can we do better? It turns out, we cannot. For example, the subset $\{1, 3, 4, 6, 7, 9\}$ contains the progression $(1, 4, 7)$. And, whatever 6-element subset of $\{1, 2, \dots, 9\}$ we choose, it will always contain such a progression—try it!³

In general, the cardinality (that is, the size) of the largest subset of $\{1, 2, \dots, N\}$ with no 3APs is denoted $r_3(N)$. For example, $r_3(9) = 5$. Another example is $r_3(30) = 12$, illustrated by the set $\{1, 3, 4, 8, 9, 11, 20, 22, 23, 27, 28, 30\}$, and in fact this is the *unique* 12-element subset of $\{1, 2, \dots, 30\}$ with no 3APs.

Let A be any counter-example to the Erdős–Turán conjecture, that is, a set with upper density greater than some $\delta > 0$ containing no 3APs. Then (5.7) implies that $|A(N)| \geq \delta N$ for infinitely many values of N . Because A contains no 3APs, $r_3(N) \geq |A(N)|$ for all N . Hence, $r_3(N)/N \geq |A(N)|/N \geq \delta$ infinitely often. Hence, if we could prove that $\lim_{N \rightarrow \infty} (r_3(N)/N) = 0$, we would rule out any such counter-example, proving the Erdős–Turán conjecture.

²For any sequence a_N , $\limsup_{N \rightarrow \infty} a_N := \lim_{N \rightarrow \infty} (\sup_{m \geq N} a_m)$.

³If you would like to see a proof, here it is. If A contains 5, it can contain 4 or 6 but not both, and similarly 3 or 7, 2 or 8, and 1 or 9, so no more than 5 elements in total. So we can assume that $5 \notin A$. Because A can have at most 2 elements out of $\{1, 2, 3\}$, and at most 2 out of $\{7, 8, 9\}$, it may have 6 elements only if it contains 4 and 6. But then it excludes 2 and 8, hence it must be $\{1, 3, 4, 6, 7, 9\}$, which is also impossible due to the progression $(1, 4, 7)$, a contradiction.

Roth's Theorem

A famous theorem of Roth [326] proves the Erdős–Turán conjecture, and in fact more. It states that there exists a constant C such that

$$r_3(N) \leq C \frac{N}{\ln \ln N}, \quad \forall N.$$

In particular,

$$\lim_{N \rightarrow \infty} \frac{r_3(N)}{N} \leq \lim_{N \rightarrow \infty} \frac{C}{\ln \ln N} = 0,$$

and the Erdős–Turán conjecture follows.

In fact, the fact that any set A with positive upper density contains a 3AP immediately implies that it contains *infinitely many* 3APs. Indeed, if there existed a set A with $\bar{\delta}(A) > 0$ containing only M 3APs, then we could just remove these $3M$ elements from A and obtain a new set A' with no 3APs, and still with positive upper density.

An Improvement on Roth's Estimate

In 1999, Bourgain [73], improving Roth's estimate, showed that

$$r_3(N) \leq C' N \sqrt{\frac{\ln \ln N}{\ln N}}, \quad \forall N,$$

for some constant C' , and this was the best known bound in 2005. However, this estimate is believed to be far from optimal. For a lower bound, Behrend [36] showed in 1946 that

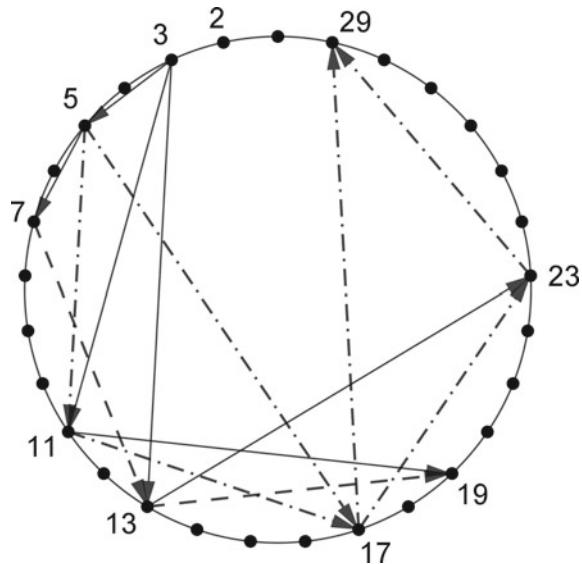
$$r_3(N) \geq N \cdot e^{-C'' \sqrt{\ln N}}$$

for some constant C'' . The difference between upper and lower bounds is huge. Let us ignore the constants for a moment, and put $C' = C'' = 1$. Then for $N = e^{100}$ the upper bound implies that $\frac{r_3(N)}{N} \leq 0.21$, while the lower bound implies only that $\frac{r_3(N)}{N} \geq 0.000045$.

3-Term Arithmetic Progressions in Primes

In particular, Bourgain's estimate tells us nothing about 3APs in some very interesting sequences such as primes. Can we form a 3AP with only prime numbers? Surely we can, for example $(3, 5, 7)$. All 3APs formed from primes less than 30 are depicted in Fig. 5.5. Are there infinitely many of them? This is a trickier question. If the set of primes had positive upper density, this would follow from Roth's theorem, but the

Fig. 5.5 Some 3-term arithmetic progressions in the primes



density of the primes is 0. In fact, the famous prime number theorem states that there are approximately $\frac{N}{\ln N}$ prime numbers not exceeding N . If we could prove that $r_3(N)$ is substantially less than $\frac{N}{\ln N}$, this would help, but, because $\frac{N}{\ln N} < C'N\sqrt{\frac{\ln \ln N}{\ln N}}$ for sufficiently large N , even Bourgain's estimate is not strong enough to be useful in this context.

However, this only means that we have no general result stating that *any* set as dense as the primes contains a 3AP. Intuitively, it should be much easier to analyse a specific sequence, like primes, than an *arbitrary* sequence with (approximately) the same number of elements. This is indeed the case. Long before Roth, van der Corput [386] proved in 1939 that the primes do indeed contain infinitely many 3APs.

A Common Generalization of the Theorems of Roth and van der Corput

Of course, we can go further and ask if there are infinitely many 3APs formed only with primes with last digit 7 (because, with the only exception of 2, the last digits of primes can be 1, 3, 7, or 9, this is about 25% of all primes), or only with primes ending with 777, or, more generally, let A be any set containing $\delta\%$ of all primes for any fixed $\delta > 0$, can we always find a 3AP which is a subset of A ?

The following theorem of Ben Green [174] gives a positive answer to all these questions.

Theorem 5.5 *Every subset of primes of positive upper density contains a 3AP.*

Of course, because the density of the primes is 0, the density (and upper density) of any subset of the primes is 0 as well. However, for a subset A of the primes, we can define an upper density *relative to the primes*, that is,

$$\bar{\delta}_P(A) := \limsup_{N \rightarrow \infty} \frac{|A(N)|}{|P(N)|},$$

where $|P(N)|$ is the number of primes not exceeding N . Theorem 5.5 states that if $\bar{\delta}_P(A) > 0$, then A must contain a 3AP. Because any finite number of 3APs can be removed from A without changing $\bar{\delta}_P(A)$, Theorem 5.5 in fact implies that A contains infinitely many 3APs.

Further Research

After Theorem 5.5 had been published, there was further progress in bounding $r_3(N)$ from above. Improving on Bourgain's estimate, Sanders [331] proved that $r_3(N) \leq \frac{N}{\ln N} (\ln \ln N)^6$, and Bloom [60] improved this to $r_3(N) \leq \frac{N}{\ln N} (\ln \ln N)^4$. If someone could prove that $r_3(N) \leq f(N) \frac{N}{\ln N}$, where $f(N)$ is a function which goes to 0 as N goes to infinity, this would give an alternative proof of Theorem 5.5. However, such an estimate is still out of reach.

Also, Green and Tao [175] extended Theorem 5.5 to longer arithmetic progressions, see Sect. 8.2 below for details.

Reference

B. Green, Roth's theorem in the primes, *Annals of Mathematics* **161**-3, (2005), 1609–1636.

5.6 Every Separable Infinite-Dimensional Banach Space Has Infinite Diameter

Vectors and Their Coordinates

Vectors in the plane are line segments with a definite direction, which connect an initial point O with a terminal point A , and are usually denoted \mathbf{OA} or just \mathbf{a} . The *length* of a vector, denoted $\|\mathbf{OA}\|$ or just $\|\mathbf{a}\|$, is the length of the corresponding line segment. Two vectors are considered to be equal if they have the same length and direction. For example, if $ABCD$ is any parallelogram, then $\mathbf{AB} = \mathbf{DC}$, because the opposite sides of a parallelogram have equal length and are parallel to each other, which indicates the same direction. Vectors can be added using the triangle rule, that is, $\mathbf{AB} + \mathbf{BC} = \mathbf{AC}$ for any triangle ABC , and can also be multiplied by a non-negative constant: for any vector \mathbf{u} and constant $\lambda \geq 0$, $\lambda\mathbf{u}$ is the vector with the same direction as \mathbf{u} and length $\lambda\|\mathbf{u}\|$. We can then define $-\mathbf{u}$ as a vector of the same

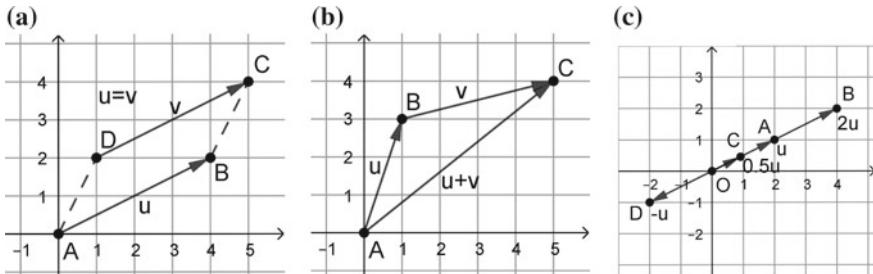


Fig. 5.6 Illustration of vector operations

length but in the opposite direction, and multiplication of \mathbf{u} by a constant $\lambda < 0$ as $-(|\lambda| \mathbf{u})$, see Fig. 5.6.

It is convenient to represent vectors using coordinates. If O is the center of the coordinate plane, the coordinates of the vector \mathbf{OA} are just the coordinates (x, y) of the terminal point A . It is easy to check that the sum of two vectors with coordinates (x_1, y_1) and (x_2, y_2) is the vector with coordinates $(x_1 + x_2, y_1 + y_2)$, and the product of a vector with coordinates (x, y) and constant λ is the vector with coordinates $(\lambda x, \lambda y)$. The length of a vector with coordinates (x, y) is $\sqrt{x^2 + y^2}$. Now we can forget about geometry and think about vectors as just ordered pairs of real numbers. Many geometric problems can in principle be solved using coordinates.

The Connection Between Vectors in the Plane and Linear Polynomials

In mathematics, there are many other “objects” which can be added or multiplied by a constant, for example, we can do these operations for functions $f : \mathbb{R} \rightarrow \mathbb{R}$, and maybe the simplest examples of functions are linear polynomials of the form $f(x) = ax + b$ for some $a, b \in \mathbb{R}$. Of course, we can add such polynomials, $(ax + b) + (cx + d) = (a + c)x + (b + d)$, and multiply by a constant, $\lambda(ax + b) = (\lambda a)x + (\lambda b)$. We can also define how “big” a linear polynomial is, for example, we can say that the “size” $\|ax + b\|$ of the polynomial $ax + b$ is $|a| + |b|$.

Similarly to vectors, linear polynomials can be very naturally associated with ordered pairs of real numbers, namely, the polynomial $ax + b$ can be associated with the pair (a, b) . Then the addition rule $(a, b) + (c, d) = (a + c, b + d)$, and the multiplication by constant rule $\lambda(a, b) = (\lambda a, \lambda b)$ are *exactly* the same as for vectors! Hence, if we have solved any problems with vectors which only involve addition and multiplication by constants, we have automatically solved the same problem about polynomials. For example, there is a theorem that *any* vector \mathbf{c} can be represented as a linear combination of vectors \mathbf{a} and \mathbf{b} if and only if \mathbf{a} and \mathbf{b} are not collinear (that is, there is no constant λ such that $\mathbf{a} = \lambda\mathbf{b}$). If we prove this theorem, we automatically prove that any linear polynomial $f(x)$ can be represented as a linear combination of

linear polynomials $g(x)$ and $h(x)$ if and only if $g(x)$ and $h(x)$ are not multiples of each other. This is just the same statement in a different language.

Vector Spaces and Isomorphisms Between Them

In order not to re-prove the same theorem thousands of times in the language of ordered pairs of numbers, vectors, polynomials, etc., it is convenient to introduce some *general* theory which covers all such cases at once. A (real) vector space V is a collection of any objects (pairs of numbers, vectors, polynomials, anything, but we will call them “vectors” for concreteness), which can be added together and multiplied by real numbers, such that the usual rules hold. For example, $x + y = y + x$, $x + (y + z) = (x + y) + z$, there exists a $\mathbf{0} \in V$ such that $\mathbf{0} + x = x \ \forall x$ etc., see Sect. 2.11 for the full list of rules. Two linear spaces V and V' are called *isomorphic* if there exists a one-to-one correspondence $f : V \rightarrow V'$ between their elements, which preserves addition and multiplication by constants, that is, $f(x + y) = f(x) + f(y)$, $\forall x, y \in V$, $f(\lambda x) = \lambda f(x)$, $\forall x \in V, \lambda \in \mathbb{R}$. The function f is called an *isomorphism*. The discussion above implies that the set V of all vectors in the plane and the set V' of all linear polynomials in one variable are isomorphic vector spaces.

Normed Vector Spaces

A vector space is called *normed* if we can associate with every $x \in V$ its length, or norm $\|x\|$, such that (i) $\|\mathbf{0}\| = 0$ but $\|x\| > 0$ if $x \neq \mathbf{0}$, (ii) $\|\lambda x\| = |\lambda| \cdot \|x\|$, and (iii) $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality). Both the examples described above are normed vector spaces, the triangle inequality for linear polynomials with norm $\|ax + b\| := |a| + |b|$ reduces to $|a + c| + |b + d| \leq (|a| + |b|) + (|c| + |d|)$ as can be easily verified, all the other properties trivially hold. However, while the two spaces are identical as vector spaces, the norms in these spaces are not the same: while for a plane vector (a, b) the norm is $\sqrt{a^2 + b^2}$, the norm of the corresponding polynomial $ax + b$ is $|a| + |b|$. Is this a big difference?

Well, if $a = b \neq 0$, then $|a| + |b| = 2|a|$, while $\sqrt{a^2 + b^2} = \sqrt{2}|a|$, so the polynomial norm is bigger by a factor of $\sqrt{2}$. Can it be worse? It turns out, it cannot. For any real a, b the inequality $(|a| - |b|)^2 \geq 0$ implies $a^2 + b^2 \geq 2|a||b|$, or $2(a^2 + b^2) \geq (|a| + |b|)^2$, which implies $\sqrt{2}\sqrt{a^2 + b^2} \geq |a| + |b|$. In the opposite direction, $2|a||b| \geq 0$ implies $(|a| + |b|)^2 \geq (a^2 + b^2)$, or $|a| + |b| \geq \sqrt{a^2 + b^2}$, hence the polynomial norm cannot be smaller. So, the norms are “almost” the same, up to a factor of $\sqrt{2}$.

The Banach–Mazur Distance Between Normed Vector Spaces

In general, we say that *normed* vector spaces V and V' are isomorphic if there exist a isomorphism $f : V \rightarrow V'$ and constants $m > 0$ and $M > 0$ such that $m\|x\|_V \leq \|f(x)\|_{V'} \leq M\|x\|_V$, $\forall x \in V$. If the constants m and M are chosen to be best possible, then we denote by $d_f(V, V')$ the ratio M/m , which indicates how much the spaces differ in terms of their norms. In our example, $d_f(V, V') = \sqrt{2}$.

In general, there may be many isomorphisms f between V and V' , and it is natural to select one with $d_f(V, V')$ as small as possible. This smallest possible value of $d_f(V, V')$ is usually denoted as just $d(V, V')$, and is called the *multiplicative Banach–Mazur distance* between V and V' . In our example, there is no isomorphism f with better ratio M/m , so in fact $d(V, V') = \sqrt{2}$.

The Diameter of a Normed Vector Space

In fact, there may be many normed vector spaces isomorphic to a given one. For example, the set of ordered pairs of real numbers (a, b) with norm $\|(a, b)\| := \max\{|a|, |b|\}$ is another normed vector space V'' . It is easy to check that $\max\{|a|, |b|\} \leq \sqrt{a^2 + b^2} \leq \sqrt{2} \max\{|a|, |b|\}$, hence $d(V, V'') \leq \sqrt{2}$. Also, the inequality $\max\{|a|, |b|\} \leq |a| + |b| \leq 2 \max\{|a|, |b|\}$ implies that $d(V', V'') \leq 2$.

In a similar way, we can introduce infinitely many different norms in the same space of ordered pairs, and it would be convenient to have an upper bound on how much they can differ from one another. For any normed vector space V , its *diameter* $D(V)$ is the *maximal* possible value of $d(V', V'')$ for any vector spaces V' and V'' isomorphic to V . There is a theorem that for our space V of ordered pairs $D(V) \leq 2$, so whatever norms we introduce they will never differ by a factor of more than 2.

More generally, for the space V_n of ordered n -tuples (x_1, x_2, \dots, x_n) with coordinate-wise addition and multiplication by a constant (this is called an n -dimensional vector space), it is known that $cn \leq D(V_n) \leq n$ for some constant c independent of n .

The Diameter of Infinite-Dimensional Spaces

However, there are vector spaces (for example, the space of all continuous functions on $[0, 1]$, or the space of all convergent sequences of real numbers $(x_1, x_2, \dots, x_n, \dots)$) whose elements cannot be described by any finite number of parameters/coordinates. Such spaces are called infinite-dimensional. Motivated by the inequality $cn \leq D(V_n)$ for n -dimensional spaces, it is natural to conjecture that the diameter of any infinite-dimensional space should be infinite. The following theorem of Johnson and Odell [211] resolves this important and difficult conjecture for the case of separable Banach spaces.

A normed vector space V is called a *Banach space* if for every sequence of vectors $x_1, x_2, \dots, x_n, \dots$ such that $\sum_{n=1}^{\infty} \|x_n\| < \infty$, there exists an $x \in V$ such that $\sum_{n=1}^{\infty} x_n = x$, see Sect. 2.11 for a detailed discussion and examples. A Banach space V is called *separable* if it contains a sequence $x_1, x_2, \dots, x_n, \dots$ such that for any $x \in V$ and any $\varepsilon > 0$ there exists an n such that $\|x - x_n\| < \varepsilon$. For example, the space of all vectors in the plane is a separable space, with x_n being, for example, the sequence of all vectors with rational coordinates.

Theorem 5.6 *If V is a separable infinite-dimensional Banach space, then $D(V) = \infty$.*

Reference

W. Johnson and E. Odell, The diameter of the isomorphism class of a Banach space, *Annals of Mathematics* **162**-1, (2005), 423–437.

5.7 The NP-Hardness of the 1.36...-Approximation to the Minimum Vertex Cover

Policemen in a City and the Minimal Vertex Cover Problem

Assume that all streets in a large city need to be controlled by the police during some important event, and that any street should be visible to at least one policeman. The policemen should not move, each one should stay at one point during the event. It is natural to put policemen at crossroads, so that each policeman can see several streets at once. For simplicity, let us assume that each street connects exactly two crossroads, and there are no crossroads in between (if there is a long street from crossroad A to C passing through crossroad B on the way, we will count this as *two* streets: from A to B and from B to C ; if there is a street from crossroad A to a dead end, we will count this dead end as a “crossroad” with only one street leading from it). We assume that, from each crossroad, a policeman can observe all the streets connecting this crossroad to the next ones, but no further. The natural question is: what is the minimal number of policemen necessary to observe all the streets?

The standard mathematical model for this type of problem is a *graph*, that is, a collection of “vertices”, some of which are connected by “edges”. In our example, vertices are crossroads and edges are streets. A *vertex cover* in any graph G is a set S of vertices such that for each edge AB at least one of the vertices A, B belongs to S . Our problem is then to find a minimal vertex cover, that is, a vertex cover with the minimal number of vertices.

The Simplest Algorithm

For example, in a graph with three vertices A, B, C , each two of which are connected by an edge, the minimal vertex cover has size 2, for example, we can “put policemen” at vertices A and B . This is clearly minimal, because one policeman, say, at vertex A , would leave the “road” BC unobserved. For a similar reason, if the graph G consists of n vertices, with each pair being connected by an edge, the minimal vertex cover has size $n - 1$. However, the situation can be much better: for example, in a graph with vertices A, B, C, D, E and edges AB, AC, AD , and AE , the minimal vertex cover has size 1—only one policeman at vertex A suffices.

To solve the problem in general, we need an *algorithm*, which, given an arbitrary graph G , outputs the minimal vertex cover in it.

One possible algorithm is just to try all the possibilities. For any vertex $A \in G$ we can either include it in the cover S or not. Hence, for the graph with n vertices, we have 2^n possibilities how to form S . For each variant, we can easily check if S is a vertex cover or not, and if so, record the size of S . Then the lowest recorded size will be the solution to our problem.

Is There a Faster Algorithm?

The issue with this algorithm is that it takes too long to run. For a city with $n = 100$ crossroads, which is not a big city at all, the total number of possibilities would be $2^{100} > 10^{30}$. If our computer could analyse 10^{10} of them per second, it would still require longer than 10^{20} s, which is more than a trillion years.

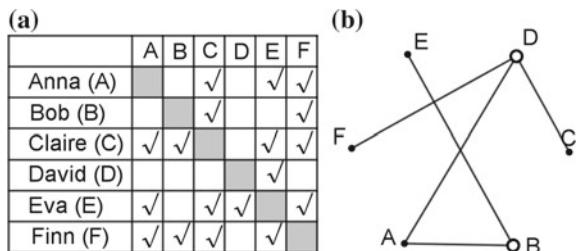
In theoretical computer science, it is common to call an algorithm *efficient* (or *polynomial*), if it runs in a number of operations bounded by some polynomial P of the size of input. So, for a graph with n vertices, we are looking for an algorithm which can enumerate n^2, n^3 , or even n^5 possibilities, but not 2^n .

So far, no-one has found a polynomial algorithm for the minimum vertex cover problem. Can we prove that no such algorithm exists? The answer is also negative, but imagine how difficult such a proof would be: there are infinitely many possible algorithmic techniques, and we would need to somehow prove that *none* of them could possibly work. Currently, mathematicians are very far from any proof like this.

Using an Algorithm for One Problem to Solve Another One

It is much easier to prove a statement like “if we can efficiently solve problem A, then we could also efficiently solve problem B”. For example, consider a group of people, each two of which are either friends or not, and assume that we need to select a maximal subset S of them such that *all* selected people are friends with each other. Let us represent people as vertices of a graph G , and connect any two vertices by an edge if and only if the corresponding people are *not* friends, see Fig. 5.7. Then the

Fig. 5.7 An example: a minimal vertex cover is $S' = \{B, D\}$, hence $S = \{A, C, E, F\}$



problem reduces to selecting a maximal subset S of vertices of G such that there is no edge connecting vertices $a \in S$ and $b \in S$.

Now, assume that we have an efficient algorithm which solves the minimum vertex cover problem. We can apply the algorithm to the graph G and let S' be a minimal vertex cover. Then, let S be the set of all vertices of G which are *not* in S' . If there were an edge connecting any $a \in S$ and $b \in S$, then this edge would be “uncovered” by the “policeman” in S' , a contradiction. Hence, there are no such edges, meaning that all people in the set S are friends with each other. Success! We have used an algorithm for one problem to solve another one!

NP-Hard Problems

There is a big class of problems, called NP , such that, if the solution is already found, there exists a polynomial algorithm which can at least *check* that the solution is correct. For example, the problem which, given a graph G and natural number k , asks if there is a vertex cover of size at most k , clearly belongs to class NP : if someone presented us with a set S of vertices which he claims solves the problem, it would be easy to verify that (i) the size of S is indeed at most k , and (ii) S is indeed a vertex cover. The issue is that to *find* a solution seems to be much harder than to *check* it!

However, Cook and Levin [101] proved in 1971 that there are some problems such that, if you could find a polynomial algorithm for them, you could use it to build a polynomial algorithm for *any* problem which belongs to class NP ! Such problems are called NP -hard problems. Because no-one really believes that it is possible to efficiently solve all problems in NP , the fact that a problem is NP -hard is considered to be very strong evidence that there is no polynomial algorithm for solving it. And one of the first problems which was shown to be NP -hard is the minimum vertex cover problem.

An Approximate Solution to the Vertex Cover Problem

If it is impossible to develop an efficient algorithm to solve a problem exactly, maybe we can at least efficiently find an approximate solution? For example, here is a very simple procedure. First, select any street, call it S_1 , and put policemen at *both* crossroads which are at the ends of the selected street. Then select any street that these two policemen cannot see, call it S_2 , and again put policemen at both ends of it. Continue this procedure until, after k steps, no streets are left uncovered. For this procedure, we need $2k$ policemen in total.

How close is this solution to an optimal one? Well, none of the streets S_1, S_2, \dots, S_k share a crossroad, so any policeman can observe at most one of them, and therefore at least k of them are really needed. Hence, our solution is within a factor of at most 2 from being optimal.

On the Possibility of a Better Approximation

The described procedure looks highly non-optimal. First, we select our streets S_1, S_2, \dots, S_k in an arbitrary way, not using some “smart” procedure. Second, we put a policeman at *both* ends of these streets, so that S_1, S_2, \dots, S_k are observed twice. It seems that it should be easy to improve this algorithm and efficiently find a vertex cover within a factor of, say, 1.99 from the optimal one. However, somewhat surprisingly, no-one has been able to achieve this, and some people conjecture that it is impossible: that is, even the problem of finding an approximate solution to the vertex cover with a factor of $2 - \varepsilon$ for any $\varepsilon > 0$ may be NP-hard.

For a long time, researchers had no methods to prove the NP-hardness of approximation problems. Then such a tool, called the “PCP Theorem” [27], was developed in the 1990s, which made it possible to prove that there exists a $\delta > 0$ such that a minimum vertex cover is *NP*-hard to approximate within a factor of $1 + \delta$. Håstad [192] in 2001 showed that it is possible to select $\delta = 1/6$ (that is, a minimum vertex cover is *NP*-hard to approximate within a factor smaller than $7/6$), but it was clear that substantially new ideas are required to improve this further. Such new ideas were developed in 2005 by Irit Dinur and Samuel Safra [118], leading to the following theorem.

Theorem 5.7 *Given a graph G , it is *NP*-hard to approximate a minimum vertex cover to within any factor smaller than $10\sqrt{5} - 21 = 1.3606\dots$.*

Further Research

To prove Theorem 5.7, the authors introduced several new powerful techniques. In particular, they demonstrated that a mathematical method called “Fourier analysis” may be useful in this context, and many further papers in the area are using this idea. They also (implicitly) used the notion of so-called “2-to-2 games”. Motivated by

this, Subhash Knot [224] introduced the so-called “2-to-2 Games Conjecture”, and showed that it would imply that a minimum vertex cover is NP -hard to approximate to within any factor smaller than $\sqrt{2} = 1.4142\dots$, as well as a number of other important and interesting consequences. This conjecture was later proved in [226] based on a series of papers, to which Dinur and Safra also contributed.

Reference

I. Dinur and S. Safra, On the hardness of approximating minimum vertex cover, *Annals of Mathematics* **162**-1, (2005), 439–485.

5.8 Embedding Large Subsets of Finite Metric Spaces into Euclidean Space

The Standard Way to Measure Distance

How would you measure the distance between two cities? The most straightforward way is to model cities as points A and B in the plane, and then the distance between them, denoted $\rho(A, B)$, is just the length $|AB|$ of the line segment joining the cities. In practice, we can measure this length on a map, and then scale appropriately, or, alternatively, learn the coordinates of the cities, say, their longitudes and latitudes, and substitute them into a formula. Mathematically, the distance between points A with coordinates (x_A, y_A) and B with coordinates (x_B, y_B) in the plane is given by

$$\rho(A, B) := |AB| = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}.$$

For example, if $A = (1, 0)$ and $B = (0, 1)$, then $|AB| = \sqrt{2}$, see Fig. 5.8a.

The distance $\rho(A, B) := |AB|$ satisfies some obvious properties. For example, for every A, B : (i) $\rho(A, B) \geq 0$, (ii) $\rho(A, B) = 0$ if and only if $A = B$, (iii) $\rho(A, B) = \rho(B, A)$, and (iv) the triangle inequality:

$$\rho(A, B) + \rho(B, C) \geq \rho(A, C), \quad \forall A, B, C. \tag{5.8}$$

Some Non-standard Ways to Measure Distance

The above “straight-line” approach to measuring the distance between cities may not be very practical. For example, imagine that you have a car, and that the above “straight-line” distance between A and B is 5 km, but there is a large forest in the way, and the shortest highway from A and B goes around the forest and has length 30 km. Then, for you, the “practical” distance between cities is rather 30 km, not 5. In general, for a person with a car, the “practical” distance $\rho(A, B)$ between any cities A, B can be defined as the shortest path from A to B through highways. For

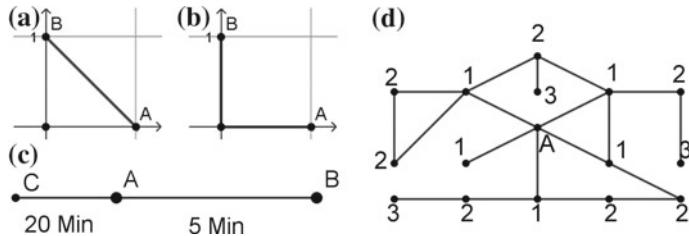


Fig. 5.8 Various ways to measure distance

example, imagine that “highways” on the coordinate plane are all parallel to its axes. Then for $A = (1, 0)$ and $B = (0, 1)$, the shortest path goes through $(0, 0)$ (or $(1, 1)$), and $\rho(A, B) = 2$, see Fig. 5.8b. Interestingly, with this definition, ρ still satisfies properties (i)–(iv) listed above.

A person who uses public transport could measure the “distance” between any points A and B in minutes, how long it takes to get from A to B . For example, if there is a direct, fast, and frequent bus from your home to your job, you will feel that your job is pretty “close”, just “5 minutes away”, while a shop in your area, but in a different direction, with no bus connection, may require you to walk 20 minutes to it, and, for you, it is “further away” than your job, see Fig. 5.8c.

Measuring the “Distance” Between People

How to measure the distance between people? Of course, we can observe their physical location, and measure the distance in the usual way. However, did you ever feel that a friend who is currently living in another country is somehow “close” to you, because you chat every day on social media, and can call each other using Skype at any moment you want? In contrast, the chancellor of your University, or a football star playing in your local team, while physically close, seems to be “far away”, living in a “different world”, and it is hard to imagine you having a cup of coffee together. We can call “friends” any two people who can easily call each other if needed, and define all your friends to be at “distance” 1 to you, then friends of your friends at “distance” 2, their friends at “distance” 3, and so on. Assuming the world is “connected”, that is, there is a chain of friends between any two people A and B , we can define $\rho(A, B)$ to be the shortest “length” of such a chain. Once again, this “distance” ρ satisfies all the properties (i)–(iv) listed above. In Fig. 5.8d, people are depicted as points (vertices), and any two points are connected by a line (edge) if and only if the corresponding people are friends. The figure shows the “distance” from the person labelled A to every other person.

Metric Spaces and Their Graphical Representation

There are various kinds of examples where we need to define the “distance” between some “objects” in many different ways, and mathematicians need a theory which can study all these situations at once. This is the theory of metric spaces. A *metric space* is any set X together with a function which assigns to every pair $A, B \in X$ a real number $\rho(A, B)$, called a *distance*, which satisfies the properties (i)–(iv) listed above. If X is a finite set, then the metric space is called finite. For example, the set of all people in the world with the distance defined above is a finite metric space.

It would be convenient if you could draw a picture in the plane (or on a computer screen), such that each person is represented as a point, and any two people A and B with small $\rho(A, B)$ are displayed on the screen close to each other, while any two people A and B with large $\rho(A, B)$ are displayed far from each other. Ideally, the (standard Euclidean) distance between points representing A and B should be exactly $\rho(A, B)$. You could then virtually observe how many people are, say, at distance at most 3 from you, or how “far” famous TV stars, the British Queen, the USA president, etc., are.

The Impossibility of an Exact Representation

Unfortunately, this exact representation is rarely possible. For example, three people who are all friends with each other could be easily visualized as an equilateral triangle with side 1. But what about four people who are all friends? We would need to represent them as 4 points A, B, C, D in the plane such that the distance between any 2 of them is exactly one. You can easily check that this is impossible. Such 4 points form a regular tetrahedron with side 1 in 3-dimensional space, but there is no way to draw them in 2 dimensions.

You may agree that the tetrahedron representation is still acceptable, and, more generally, allow representations in d -dimensional space \mathbb{R}^d , where every point is described by d coordinates, and the distance between any two points with coordinates (x_1, x_2, \dots, x_d) and (y_1, y_2, \dots, y_d) is given by $\sqrt{(x_1 - y_1)^2 + \dots + (x_d - y_d)^2}$. Unfortunately, this does not help much. For example, imagine four people A, B, C, D such that A is a friend to everyone, but B, C, D do not know each other. Then $\rho(A, B) = \rho(A, C) = \rho(A, D) = 1$ but $\rho(B, C) = \rho(C, D) = \rho(D, B) = 2$. Can we draw 4 points with such distances? No! If we could, B, C , and D would form a equilateral triangle with side 2. Because $\rho(A, B) = \rho(A, C) = 1$, A should be a midpoint of side BC . But, on the other hand, $\rho(A, B) = \rho(A, D) = 1$, hence A should also be a midpoint of BD , a contradiction.

Approximate Solutions: Embeddings with Distortion

Ok, if an exact drawing is impossible, can we find an approximate one? After all, if friends were displayed at distance 0.98 instead of 1, or “friends of the friends”

at distance 2.05 instead of 2, the picture would still be very useful. Formally, a “drawing”, or *embedding*, is just a function f from a metric space X to \mathbb{R}^d . If there are constants $m, M > 0$ such that

$$m\rho(A, B) \leq |f(A)f(B)| \leq M\rho(A, B), \quad \forall A, B \in X,$$

where $|f(A)f(B)|$ is the usual length of the line segment with endpoints $f(A)$ and $f(B)$, we will say that the *distortion* of the embedding f into \mathbb{R}^d is at most M/m . There is a theorem stating that any n -element metric space can be embedded into \mathbb{R}^d with distortion at most $C \ln n$ (where the dimension d depends on n but the constant C does not). However, in our example, with n being billions of people in the world, $\ln n$ would be more than 20, so friends could be displayed at distance 20, while people with shortest chain of friends 20 could be displayed at distance 1, which makes the “picture” totally useless. We would really need a picture with *much* smaller distortion. However, there is a theorem that tells us, if we insist on embedding a whole metric space (in our example, really all the people in the world), the bound $C \ln n$ cannot be significantly improved.

Embedding a Subset with Better Distortion

Can we achieve a much better distortion if we agree to display not all people in the world, but only a significant part of them? It turns out we can, and this is what the following theorem of Bartal et al. [35] is about.

Theorem 5.8 *There exists an absolute constant $C > 0$ such that for every $\alpha > 1$, every n -point metric space has a subset of size $n^{1-C\frac{\ln(2\alpha)}{\alpha}}$ which can be embedded into \mathbb{R}^d with distortion at most α .*

Here, the dimension d depends on n but the constant C does not. Assuming, for the sake of illustration, that $C = 1$, and taking, say, $\alpha = 3$, we can conclude that we can “display” a subset of size about $n^{1-\frac{\ln 6}{3}} \approx n^{0.4}$ with distortion at most 3, which is a reasonable approximation for many applications.

Using the tools developed to prove Theorem 5.8, the last two authors achieved an even better result in a subsequent work [270]: they showed that we can take an even larger subset of size $n^{1-\frac{C}{\alpha}}$ such that the conclusion of Theorem 5.8 remains correct. This last result is optimal up to the value of the constant C .

Reference

Y. Bartal, N. Linial, M. Mendel, and A. Naor, On metric Ramsey-type phenomena, *Annals of Mathematics* **162**-2, (2005), 643–709.

5.9 On the Number of Quartic Fields with Bounded Discriminant

Numbers: Integers, Rational, Real, and Complex

There are different types of numbers needed for various applications. If you just need to count objects, it suffices to study natural numbers $0, 1, 2, \dots$. If you would also like to be able to perform basic mathematics operations, you need to introduce negative numbers to be able to subtract, for example, 5 from 3, and, more generally, rational numbers, to be able to divide, for example, 2 by 7. With the set \mathbb{Q} of rational numbers it is possible to perform all basic operations (addition, multiplication, subtraction, and division), except for division by 0. Any set for which these operations can be defined, such that all the usual properties hold (for example, commutativity $a + b = b + a$ and $ab = ba$, associativity $a + (b + c) = (a + b) + c$ and $a(bc) = (ab)c$, etc.), is called a *field*. However, we need “other” numbers if we would also like to solve equations. For example, the equation $x^2 - 5 = 0$ has no rational solution: the solutions are $\pm\sqrt{5}$, where $\sqrt{5} = 2.236\dots$ can be written in the form of an infinite decimal, and the set of all such numbers $a.bcde\dots$ is called the field of real numbers \mathbb{R} . Moreover, even in \mathbb{R} we cannot solve, for example, the equation $x^2 + 1 = 0$. Let us just denote the solution by i and “adjoin” it to \mathbb{R} . To be able to perform multiplication we should then also allow numbers of the form bi , $b \in \mathbb{R}$, and, to be able to perform addition, we should also allow numbers $a + bi$, $a, b \in \mathbb{R}$. The set \mathbb{C} of all numbers in this form is called the set of *complex* numbers, and it is also a field, see Sect. 1.7 for details and more examples. Geometrically, the complex number $a + bi$ can be represented as a point with coordinates (a, b) in the plane. It turns out that any n -degree polynomial $P(x) = a_nx^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$ can be equivalently written as $P(x) = a_n(x - x_1)(x - x_2) \dots (x - x_n)$ for some complex numbers x_1, x_2, \dots, x_n , which are solutions to the equation $P(x) = 0$. Hence, we have finally added “enough” numbers to be able to solve all polynomial equations.

The Field $\mathbb{Q}(\sqrt{5})$

However, if we care only about basic operations and polynomial equations, we have in fact added “too many” numbers. While moving from \mathbb{R} to \mathbb{C} , we have added “just” the root i of equation $x^2 + 1 = 0$, and the numbers $a + bi$ needed for arithmetic operations with i . However, while moving from \mathbb{Q} to \mathbb{R} , we, for some reason, have added not only roots of equations we care about (such that $x^2 - 5 = 0$), but *all* possible infinite decimal numbers. For example, the well-known constants $\pi = 3.14159\dots$ and $e = 2.71828\dots$ are real numbers, which are *not* solutions to any polynomial equation with rational coefficients.

What if, similar to the “real-complex” transition, we just add the root of the equation we care about? For example, take the equation $x^2 - 5 = 0$, and “adjoin” its root $\gamma = \sqrt{5}$ to \mathbb{Q} . To be able to perform addition and multiplication, we should also

adjoin all numbers of the form $a + b\gamma$, $a, b \in \mathbb{Q}$. It turns out that this set (let us denote it by $\mathbb{Q}(\gamma)$) is also a field, we can perform all basic operations and always get a number of the same form! Indeed, $(a + b\gamma) \pm (c + d\gamma)$ is just $(a \pm c) + (b \pm d)\gamma$. For multiplication, $(a + b\gamma) \cdot (c + d\gamma) = ac + (ad + bc)\gamma + bd\gamma^2$. But $\gamma^2 = 5$, and this reduces to $(ac + 5bd) + (ad + bc)\gamma \in \mathbb{Q}(\gamma)$. Finally, $(a + b\gamma) \cdot (a - b\gamma) = a^2 + 5b^2$, hence $(a + b\gamma)^{-1} = (a - b\gamma)/(a^2 + 5b^2)$ also belongs to $\mathbb{Q}(\gamma)$, and so does the result of any division $(c + d\gamma)/(a + b\gamma) = (c + d\gamma)(a - b\gamma)/(a^2 + 5b^2)$.

Algebraic Integers in $\mathbb{Q}(\sqrt{5})$

Each number $a + b\gamma \in \mathbb{Q}(\gamma)$ is a solution to the equation $x^2 - 2ax + (a^2 - 5b^2) = 0$ with rational coefficients $1, -2a, a^2 - 5b^2$. If these coefficients are in fact integers, then $a + b\gamma$ is called an *algebraic integer*. This can happen if either both a, b are integers, or $a = k/2, b = m/2$ for some odd integers k and m . Equivalently, $a + b\gamma = k_1 \cdot z_1 + k_2 \cdot z_2$ for some integers k_1, k_2 where $z_1 = 1, z_2 = \frac{1+\gamma}{2}$. Geometrically, elements $a + b\gamma$ of $\mathbb{Q}(\gamma)$ can be represented as points (a, b) with rational coordinates, and then the algebraic integers form a nice regular structure in the plane which is called a lattice. In fact, the plane is fully filled by translated copies of the parallelogram with vertices $(0, 0), (1, 0), (3/2, 1/2), (1/2, 1/2)$, and the algebraic integers are presented as vertices of this tiling, see Fig. 5.9.

How many algebraic integers are there such that, say, $0 \leq a \leq T$ and $0 \leq b \leq T$ for a large constant T ? In fact, a is either an integer or “half-integer”, hence there are about $2T$ possible values for it in $[0, T]$, and, for each a , there are about T choices for b , hence the answer is approximately $2T^2$. A more elegant way to get the same answer is to notice that the area of the above parallelogram is $S = 1/2$, hence we need about $T^2/S = 2T^2$ parallelograms to fill the square $[0, T] \times [0, T]$. Thus, the

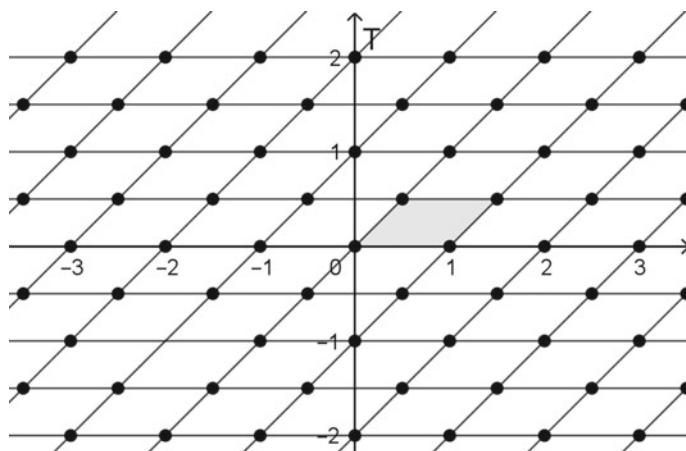


Fig. 5.9 Algebraic integers in $\mathbb{Q}(\sqrt{5})$

area S has an important interpretation as the “density” of the algebraic integers in $\mathbb{Q}(\gamma)$.

Algebraic Integers in Algebraic Number Fields

Of course, there is nothing special about the equation $x^2 - 5 = 0$. We can add a root γ of *any* polynomial equation $P(x) = 0$, where $P(x) = x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$, with $a_0, a_1, \dots, a_{n-1} \in \mathbb{Q}$. However, if $n > 2$, and γ is not a solution of any quadratic equation (for example, if $P(x) = x^3 - 5$ and $\gamma = \sqrt[3]{5}$), then γ^2 cannot be represented in the form $a + b\gamma$, $a, b \in \mathbb{Q}$, and should therefore be included separately. More generally, if γ is not a solution of any equation of degree less than n , we should include $\gamma, \gamma^2, \dots, \gamma^{n-1}$, and define $\mathbb{Q}(\gamma)$ as the set of all numbers of the form $b_0 + b_1\gamma + \dots + b_{n-1}\gamma^{n-1}$, $b_0, b_1, \dots, b_{n-1} \in \mathbb{Q}$. Such $\mathbb{Q}(\gamma)$ is called an *algebraic number field* of degree n . Again, of special interest are the elements of $\mathbb{Q}(\gamma)$ which are algebraic integers, that is, solutions of polynomial equations with *integer* coefficients. Similar to the case $\gamma = \sqrt{5}$ described above, algebraic integers in $\mathbb{Q}(\gamma)$ can be described as numbers of the form $k_1 \cdot z_1 + k_2 \cdot z_2 + \dots + k_n \cdot z_n$, where k_1, k_2, \dots, k_n are arbitrary integers, and z_1, z_2, \dots, z_n are some fixed numbers. Geometrically, elements of $\mathbb{Q}(\gamma)$ can be represented as points $(b_0, b_1, \dots, b_{n-1})$ with rational coordinates in n -dimensional space, and the algebraic integers form a regular structure (lattice) in this space. The “density” of the algebraic integers in $\mathbb{Q}(\gamma)$ is inversely proportional to the volume S of the n -dimensional parallelepiped with vertices $\sum_{i=1}^n s_i z_i$, where each s_i is either 0 or 1 (resulting in 2^n vertices in total).

For each number field $K = \mathbb{Q}(\gamma)$, we can compute a number, denoted d_K , which is called the *discriminant* of the number field K . The actual definition is a bit complicated, but it is related to the squared volume of the parallelepiped described above, and therefore tells us “how many” algebraic integers there are in K . The definition is: take z_1, z_2, \dots, z_n as above, then each $z_i \in \mathbb{Q}(\gamma)$ and therefore $z_i = \sum_{k=0}^{n-1} b_{ik} \gamma^k$ for some rational numbers b_{ik} . Next, define $c_{ij} = \sum_{k=0}^{n-1} b_{ik} \gamma_j^k$, $i = 1, \dots, n$, $j = 1, \dots, n$, where $\gamma_1, \dots, \gamma_n$ are (possibly complex) roots of the polynomial $P(x)$ we started with. Then d_K is defined as the square of the determinant⁴ of the $n \times n$ matrix with entries c_{ij} .

Counting Number Fields with Bounded Discriminant

It is far from obvious from the definition, but the discriminant d_K is always an integer. Also, for each integer m , there exist only finitely many different number fields K with

⁴The definition of the $n \times n$ determinant is also a bit complicated. Let S_n be the set of all possible permutations $\sigma = (\sigma(1), \dots, \sigma(n))$ of the set $(1, 2, \dots, n)$. For example, $(3, 1, 2)$ is a possible permutation of $(1, 2, 3)$. For general n , there are $n!$ possible permutations. Each permutation σ can be “implemented” by starting from $(1, 2, \dots, n)$ and exchange adjacent elements, e.g. $(1, 2, 3) \rightarrow (1, 3, 2) \rightarrow (3, 1, 2)$. The sign of sigma is defined as $(-1)^n$, where n is the number of steps in this sequence. The determinant of a matrix with entries c_{ij} is defined as $\sum_{\sigma \in S_n} \text{sign}(\sigma) \prod_{i=1}^n c_{\sigma(i)i}$.

discriminant m . This implies that the number $N_n(X)$ of number fields with degree n and absolute value of the discriminant at most X is also finite. Moreover, there is a conjecture that, for each fixed n , $N_n(X)$ grows linearly as a function of X for large X . More formally, the conjecture states that there is a (finite positive) constant c_n such that $\lim_{X \rightarrow \infty} N_n(X)/X = c_n$.

The conjecture is “trivial” (for experts) if $n = 2$, and has been known since 1971 if $n = 3$ [107], but no further progress had been made for decades, until the next step was made by Manjul Bhargava [51].

Theorem 5.9 *Let $N_4(X)$ be the number of number fields of degree 4 such that the absolute value of the discriminant is at most X . Then*

$$\lim_{X \rightarrow \infty} \frac{N_4(X)}{X} = c_4 \approx 0.253\dots$$

In a later work, Bhargava [53] also resolved the $n = 5$ case of this conjecture, see Sect. 10.11 for details.

Reference

M. Bhargava, The density of discriminants of quartic rings and fields, *Annals of Mathematics* **162**-2, (2005), 1031–1063.

5.10 Optimal Sphere Packing in Dimension 3

Packing Disks in the Plane: Square Packing

How many round tables can you fit inside a large restaurant, assuming that each table, with chairs and space for people around it, occupies an area in the form of a disk of radius 1 square meter? Mathematically, the question is about fitting as many disks of unit radius as possible within a given region in the plane. Because the area of the unit disk is π , the upper bound is S/π , where S is the area of the region. However, this upper bound is impossible to achieve, because disks cannot cover the whole area, there will always be some gaps between them.

In order to understand the minimal possible size of these gaps, let us try to fill the *whole* plane with unit disks as densely as possible. We can start with the disk centred at $(0, 0)$ in the coordinate plane, then add disks on the left and right as closely as possible, these are the disks with center coordinates $(-2, 0)$ and $(2, 0)$, respectively. We can continue this way and form a whole infinite row of disks, with center coordinates $(2n, 0)$, $n \in \mathbb{Z}$, where \mathbb{Z} is the set of all integers. Next, we can add the disk with center $(0, 2)$ above our initial disk $(0, 0)$, and form a similar row to the left and right from it, the coordinates are $(2n, 2)$, $n \in \mathbb{Z}$. Continuing this way indefinitely, we can “fill” the whole plane by disks with centres $(2n, 2m)$, $n, m \in \mathbb{Z}$. This is called a *square packing*.

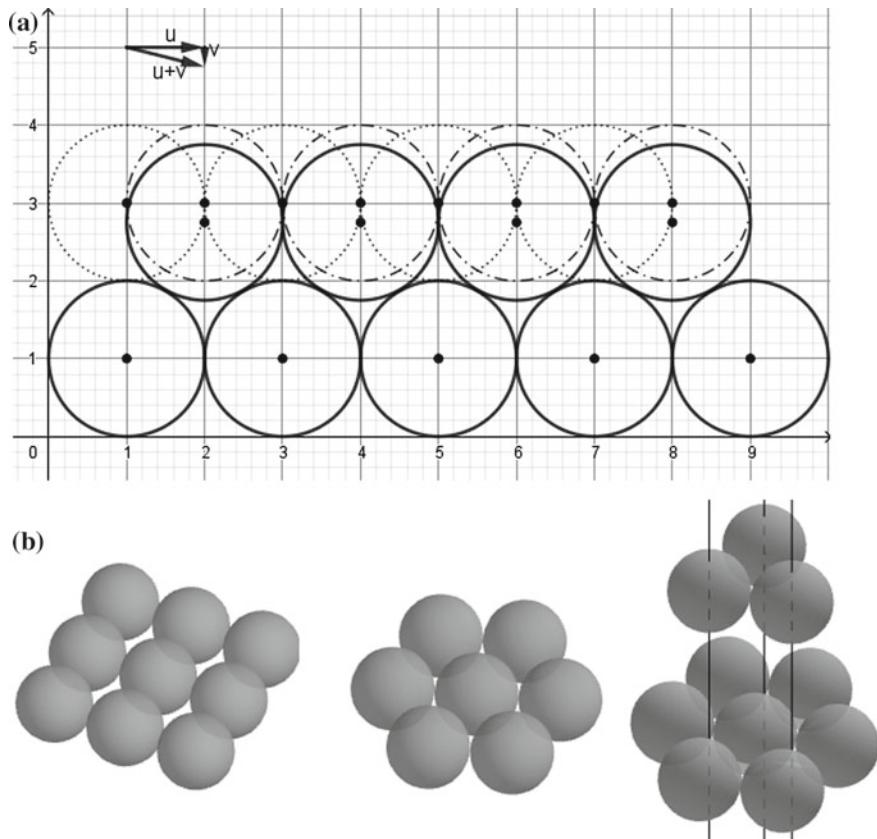


Fig. 5.10 Optimal circle and sphere packings in dimensions 2 and 3

Hexagonal Packing

Of course, there are some gaps in between these disks. Can we pack them “densely”, so that the gaps are “smaller”? It turns out, we can! In the square packing, let us fix the row of disks with centres $(2n, 0)$, $n \in \mathbb{Z}$, but shift the row above it by 1, so that the centres are now at points $(2n + 1, 2)$, $n \in \mathbb{Z}$. Now the disks in the two rows do not touch each other anymore, and the top row can be moved down, to make the packing dense, see Fig. 5.10a. The centre coordinates would be $(2n + 1, t)$, and we can move it until the distance between $(2n + 1, t)$ and $(2n, 0)$ is at least 2, that is, $\sqrt{(2n + 1 - 2n)^2 + t^2} \geq 2$, or $t \geq \sqrt{3}$. This results in a row with centre coordinates $(2n + 1, \sqrt{3})$, $n \in \mathbb{Z}$. We can continue in this way, and put the centres of the disks at points $(m, n\sqrt{3})$, where m, n are integers of the same parity (either both even or both odd). In this packing, each disk touches six other disks, whose centres form a regular hexagon, and, for this reason, it is called a *hexagonal packing*.

Estimating and Comparing Densities of Different Packings

How can we check that the hexagonal packing is “denser” than the square one? For this, consider a large square with vertex coordinates $(0, 0)$, $(0, T)$, (T, T) and $(T, 0)$ for T very large. How many disks does it contain in the square packing? Well, the center $(2n, 2m)$ is inside if $0 \leq 2n \leq T$ and $0 \leq 2m \leq T$, hence we have about $T/2$ possible choices for m and $T/2$ for n , resulting in about $T^2/4$ disks in total. The total area of all these disks is $T^2/4 \cdot \pi$, while the total area of the square is T^2 , hence the disks fill about $\pi/4 \approx 78.5\%$ of the square. Of course, if T is an even integer, there are $T/2 + 1$ choices for m and n . Also, some boundary disks may have part of their area within the square, and part outside. However, for large T , the relative effect of such “boundary conditions” is negligible.

With hexagonal packing, the center $(m, n\sqrt{3})$ lies within the square if $0 \leq m \leq T$ and $0 \leq n\sqrt{3} \leq T$. There are about $T/\sqrt{3}$ possible choices for n , and, for each n , about $T/2$ choices for m , because it should have the same parity as n . This result in about $T^2/2\sqrt{3}$ disks with area π each, which cover about $\pi/2\sqrt{3} \approx 90.7\%$ of the area of the square, much better than square packing! It has been known since the 19th century that this packing is in fact the densest possible.

Packing Spheres in 3-Dimensional Space

What if we need to find the densest possible packing in 3-dimensional space rather than in the plane? For example, how many oranges can you pack into a large box? Mathematically, oranges can be modelled as balls with unit radius, and the question is how densely they can fill the space. Formally, a *sphere packing* is a set S of points in \mathbb{R}^3 such that $\|x - y\| \geq 2$ for all distinct $x, y \in S$, where $\|\cdot\|$ denotes the usual distance, and the elements of S correspond to the centres of the spheres/balls. The 3-dimensional analogue of a square packing is a cubic packing, with ball centres at points $(2m, 2n, 2k)$ with m, n, k integers. Within a large cube with side length T we can fit about $(T/2)^3$ such centres. The volume of the unit ball is $\frac{4}{3}\pi$, hence the total volume of all balls within the cube is $\frac{1}{6}\pi T^3$. In other words, the balls occupy only $\pi/6 \approx 52\%$ of the cube volume.

Of course, this is not optimal. The cube packing consists of “layers” corresponding to $k = 0, 1, -1, 2, -2, \dots$, and, within each layer, the centres can be re-arranged using the optimal hexagonal packing in the plane rather than the square packing. A large cube with side lengths T would then contain about $T^2/2\sqrt{3}$ balls within each layer, which with $T/2$ layers results in $T^3/4\sqrt{3}$ balls, covering volume $\frac{1}{3\sqrt{3}}\pi T^3$, that is, $\pi/3\sqrt{3} \approx 60\%$ of the cube volume.

Now, similar to the 2-dimensional case, we can shift the layers so that they do not touch, and then move them closer to each other, see Fig. 5.10b. For $k = 0$, we have the balls with centres $(0, 0, 0)$, $(2, 0, 0)$, and $(1, \sqrt{3}, 0)$, forming the right triangle, so why not put a ball on the next layer with center equally far from all the vertices of this triangle? The coordinate of such a center would be $(1, \sqrt{3}/3, t)$, and the (interiors

of) balls would not intersect if $\sqrt{(1-0)^2 + (\sqrt{3}/3)^2 + (t-0)^2} \geq 2$, or $t \geq \sqrt{8/3}$. We can organise the whole layer around this ball in an optimal way, and proceed similarly. In the resulting packing, our large cube would contain $T/\sqrt{8/3}$ layers with about $T^2/2\sqrt{3}$ balls within each layer, that is, about $T^3/4\sqrt{2}$ balls in total, with total volume $\frac{1}{3\sqrt{2}}\pi T^3$, which is $\pi/3\sqrt{2} \approx 74\%$ of the cube volume. This is called a *hexagonal close packing*.

The Densest Possible Packing

Of course, instead of a large cube, we could consider a large tetrahedron, or a sphere, or any other reasonable shape, and the result would be the same, at least in the limit. For example, assume that a large ball with centre $(0, 0, 0)$ and radius R contains $|S|$ unit balls of the packing S . Then the total volume occupied by these balls is $4\pi|S|/3$, while the volume of the large ball is $4\pi R^3/3$, hence the balls in the packing covers a proportion $|S|/R^3$ of the whole volume. It can be shown that for the hexagonal close packing and large R , $|S| \approx \frac{\pi}{3\sqrt{2}}R^3$, resulting in the same $\pi/3\sqrt{2} \approx 74\%$ of the volume covered. Formally, the *density* of a sphere packing S is

$$\delta(S) := \lim_{R \rightarrow \infty} \frac{|S|}{R^3},$$

so that the density of the hexagonal close packing is $\pi/3\sqrt{2}$.

Technically, the density $\delta(S)$ may not exist for some packings S . One could imagine a packing for which the fraction $\frac{|S|}{R^3}$ is equal to a for odd R , and is equal to b for even R , with $a \neq b$, so that the limit $\lim_{R \rightarrow \infty} \frac{|S|}{R^3}$ does not exist.

In this case, we can introduce the *upper density*

$$\bar{\delta}(S) := \limsup_{R \rightarrow \infty} \frac{|S|}{R^3},$$

where $\limsup_{R \rightarrow \infty} f(R)$ is the largest number A such that $A = \lim_{k \rightarrow \infty} f(R_k)$, where R_1, R_2, \dots is a sequence such that $\lim_{k \rightarrow \infty} R_k = \infty$. The upper density $\bar{\delta}(S)$ exists for every packing S .

Kepler conjectured in the 17th century that the hexagonal close packing is the densest possible. This conjecture had been open for many centuries, until it was finally confirmed by Hales [184].

Theorem 5.10 *No packing of congruent balls in Euclidean three space has upper density greater than $\pi/3\sqrt{2}$.*

The proof of Theorem 5.10 is extremely complicated, and includes thousands of cases verified by a computer.

Reference

T. Hales, A proof of the Kepler conjecture, *Annals of Mathematics* **162**-3, (2005), 1065–1185.

5.11 The Chromatic Number of a Random Graph

From Party Organisation to Graph Colouring

Imagine you are organising a large party, and some of your guests like each other, but some do not. You would like to arrange seats for your guests in such a way that everyone shares a table only with people he/she likes. In other words, you would like to divide all guests into groups such that no two enemies are in the same group. The question is, what is the minimal number of tables/groups you would need?

For a convenient graphical illustration, you could represent every guest as a point in the plane, and then connect two points by a line if and only if the corresponding guests do not like each other. You must now colour your points in such a way that every two points connected by a line have a different colour, and the question is what is the minimal number of colours you would need.

Mathematician would call such a picture a *graph*, with points (guests) called the *vertices* of the graph, and lines (pairs of guests who are enemies) called the *edges* of the graph. A colouring as described above is called a *proper graph colouring*, and the minimal number of colours in such a colouring is called the *chromatic number* of the graph.

A Search for an Optimal Colouring

For example, if we have just four guests, whom we can label as A, B, C, D , and the pairs of enemies are (A, B) , (B, C) , (C, D) , and (D, A) , the corresponding graph can be represented in the plane as a square, and its chromatic number is 2—we can colour A and C black and B and D white, for example. So, two tables suffice, see Fig. 5.11a.

For a general graph, it is easy to check if two colours (tables) suffice. Just colour an arbitrary vertex (guest) black, all his/her enemies white, all “enemies of enemies” black, and so on, until we either get a proper colouring, or a contradiction. For example, in Fig. 5.11b we have started with vertex A , then vertices B and E are neighbours of A and are therefore white, vertices C and D are neighbours of B and E , respectively, and are therefore black, but this is a contradiction because C is connected to D .

If two colours do not suffice, we can try three colours. There is no fast and easy method for determining if a proper colouring with three colours is possible, but, if our computer is fast, and the number of guests is not too big, we could, at least

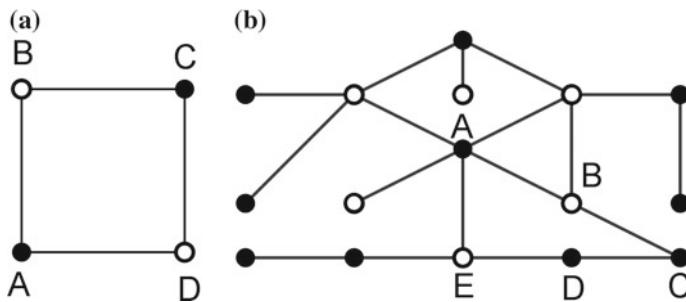


Fig. 5.11 Successful and unsuccessful 2-colourings of graphs

in principle, ask the computer to try all possible colourings. If three colours is not enough, we could try four, and so on, until success.

Allowing Solutions Which Are “Likely” to Work

However, this is possible only if we know *in advance* who likes and dislikes whom. What if we do not have the exact information in advance, and just ask everyone for the list of his/her enemies on arrival? If we have n guests, what is the number of tables which is *guaranteed* to suffice? If we formulate the question like this, the answer is very pessimistic: theoretically, it can be that all guests dislike all others, so the only solution is to prepare n tables, one for each guest. Of course, this would not be a very fun party...

To do better, why not allow us to fail, but with some very small probability. For example, if we prepare “just” $n - 1$ tables, then this is very “likely” to be sufficient. Indeed, if there are at least two guests which are not enemies, we could put them on the same table, and allocate all others to private tables. So, the only case in which we will fail is indeed if *all* pairs of guests dislike each other. But we can ignore this very special scenario as being “very unlikely”.

But what is “likely” and “unlikely”? To be able to calculate probabilities, we would need some mathematical model. And perhaps the simplest one is to assume that every pair of guests are enemies with some probability $p \in (0, 1)$, independently from other pairs. The resulting random graph is denoted $G(n, p)$. We can also assume that $p = d/n$ for some constant d , so that each person has approximately d enemies. Because there are $\frac{1}{2}n(n-1)$ pairs of guests (check!), the probability that everyone dislikes everyone is, in this model, $(d/n)^{\frac{1}{2}n(n-1)}$. If, for example, $d = 2$, then even for $n = 5$ this results in $(0.4)^{10} \approx 0.0001$, and this number quickly decreases as n increases. So, indeed, $n - 1$ colours/tables will be enough with probability very close to 1. But can we do better? What is the minimal number of colours which will be sufficient with high probability? Here, by “with high probability” (or *almost surely*), we mean “with probability tending to 1 as n goes to infinity”.

An Explicit Analysis for Small Parties

For example, let $d = 2$. If $n = 2$, then the two guests are enemies with probability $d/n = 1$, and we surely need 2 tables for them. If $n = 3$, each pair are enemies with probability $2/3$. Hence, we need 3 tables with probability $(2/3)^3 \approx 0.3$, otherwise 2 is enough (and even 1 table is enough with small probability $(1/3)^3 < 4\%$). If $n = 4$, $d/n = 1/2$, hence all pairs have a 50% chance of being enemies, and all possible graphs with 4 vertices can appear with equal chance. There are $2^6 = 64$ of them, and only in one variant (all enemies to all) would 3 colours not suffice. By direct enumeration one can count that 22 of the graphs contains a triangle, so with probability $\frac{22}{64} \approx 0.34$ exactly 3 colours are needed, while otherwise 2 colours suffice.

In these examples, one can see that in almost all cases we need either 2 or 3 colours. One can guess that this is because n was small, and, with large n , much more colours are needed. However, this intuition is wrong. A very special case of the theorem we will discuss below implies that, with $d = 2$, we would almost surely need either 2 or 3 colours!

Why Two Tables May Suffice Even for Large Parties

How can it be that, even with, say, $n = 1,000$, we have a reasonable chance that 3, or even just 2 colours suffices? After all, if we have a “triangle”—that is, three people who are pairwise enemies to each other—then surely 2 colours is not enough. And there are almost $n^3/6$ (more exactly, $n(n - 1)(n - 2)/6$) ways to select 3 people out of n . However, for each particular triple, the probability that they are all enemies is just $(2/n)^3$, so, with probability $1 - 8/n^3$ they do *not* form a triangle. Assuming that all triples are independent (which is in fact not fully correct, but the assumption is ok for rough calculation), there will be no triangles with probability about $(1 - 8/n^3)^{n^3/6}$, or $[(1 - 1/k)^k]^{4/3}$, where $k = n^3/8$. It is known from calculus that, for large k , $(1 - 1/k)^k$ is about e^{-1} , where $e \approx 2.71$, hence with reasonable probability about $e^{-4/3} \approx 0.26$ there will be no triangles at all. This of course does not imply that 2 colours suffices,⁵ but provides some intuition as to why this *may* be possible.

An Explicit Formula for the Number of Tables/Colours

In fact, it has been known since 1991 that, for any d , there is an integer k_d such that, almost surely, the chromatic number of $G(n, d/n)$ is either k_d or $k_d + 1$, see [254]. In other words, we can pre-order $k_d + 1$ tables, and can be almost sure that (i) this is enough, and (ii) this is almost optimal, in the sense that, with high probability, at most 1 table will be unused.

⁵For example, for 5 people such that exactly $(1, 2), (2, 3), (3, 4), (4, 5)$, and $(5, 1)$ are enemies, no 3 form a triangle, but 2 colours is not enough.

This fact is beautiful, but completely useless for any practical purposes, unless we know a method for actually computing k_d for any given d . This is what the following theorem of Achlioptas and Naor [1] is about.

Theorem 5.11 *For any $0 < d < \infty$, let k_d be the smallest integer k such that $d < 2k \ln k$. Then almost surely (that is, with probability that tends to 1 as $n \rightarrow \infty$), the chromatic number of $G(n, d/n)$ is either k_d or $k_d + 1$. Moreover, if also $d > (2k_d - 1) \ln k_d$, then, almost surely, the chromatic number of $G(n, d/n)$ is exactly $k_d + 1$.*

For example, if $d = 2$, then the smallest integer k such that $2 < 2k \ln k$ is $k_2 = 2$, hence, with high probability, we need either 2 or 3 colours, consistent with the calculation above. Moreover, for $d = 6$, the smallest integer k such that $6 < 2k \ln k$ is $k_6 = 3$, and also $6 > 5 \ln 3 = (2k_6 - 1) \ln k_6$, hence, with high probability, the chromatic number of $G(n, 6/n)$ is exactly 4. In other words, if, in our model, each guest has on average 6 enemies, then, with high probability, 4 tables will be sufficient for enemy-free guest allocation, while 3 tables will not. The significance of Theorem 5.11 is that it allows us to calculate the (probable) chromatic number of $G(n, d/n)$ for a given d in an explicit and easy way.

Reference

D. Achlioptas and A. Naor, The two possible values of the chromatic number of a random graph, *Annals of Mathematics* **162**-3, (2005), 1335–1351.

Chapter 6

Theorems of 2006



6.1 Sufficient Conditions for Completeness of a Set of Integers

Coin Nominals in the UK

How to select “good” coin nominals to be used in a country? To answer this question, we first need to clarify what exactly is meant by “good”. The first obvious requirement is that we should be able to pay any amount of money in a shop. In particular, if a coin with nominal 1 is absent, and we have bought an item which costs 1, we will not be able to pay for it, which is, of course, very inconvenient. On the other hand, if a coin with nominal 1 exists, we can pay any sum by using just these coins, so why do we need coins with different nominals?

Well, when buying something for 98 pence, it would be inconvenient to use 98 coins with nominal 1. For this reason, coins with nominals 2, 5, 10, 20, and 50 exist in the UK. In fact, with these coins, any sum from 1 to 100 can be paid in such a way that coins with each nominal are used at most twice. For example, $98 = 50 + 20 + 20 + 5 + 2 + 1$, only the coin with nominal 20 is used twice (Fig. 6.1).

Aiming for a Better Coin Nominal System

What if we aim for an “even better” coin nominal system, such that every sum can be paid using coins with each nominal at most *once*? To achieve this, one may just introduce coins with each possible nominal 1, 2, 3, 4, . . . , but this is inconvenient and uninteresting. For example, if we have coins 1 and 2, introducing a coin with nominal 3 is unnecessary, because $3 = 2 + 1$. Next, we need nominal 4, but not 5, 6, 7, because $5 = 4 + 1$, $6 = 4 + 2$, and $7 = 4 + 2 + 1$. Then we need 8, and, continuing this way, we conclude that we can include coins with nominals 2^k , $k = 0, 1, \dots$, and will be able to pay any sum using only distinct coins. In other words, any integer can be represented as a sum of distinct powers of 2.



Fig. 6.1 Forming 98 pence from legitimate coins

Of course, powers of 2 is not the only possible choice for such a coin system. For example, let $P = \{1, 2, 3, 5, 7, 11, \dots\}$ be the set of all prime numbers, together with the number 1. Then, once again, any integer can be represented as a sum of distinct elements of P , hence P can also be selected as a set of nominals for a “good” coin system. Indeed, for any integer n we can select the largest prime p such that $p \leq n$. By the famous Bertrand’s postulate, proved by Chebyshev in 1852, $p > n/2$, hence the remainder $r = n - p$ is less than $n/2$. We can next select the largest prime not exceeding r , and repeat the procedure until the remainder is 0 or 1.

Of course, the set P is a bit strange, because 1 is not a prime number. However, if we allow only coins with prime nominals, then we cannot pay for a purchase which costs 1, and also $2 + 2$ is the only way to represent 4, while $3 + 3$ is the only way to represent 6, contrary to our requirement that coins with each nominal should be used at most once. However, Richert [321] proved in 1949 that these are the only exceptions, and every integer greater than 6 can be represented as a sum of *distinct* primes. For example, $7 = 7$, $8 = 5 + 3$, $9 = 7 + 2$, and so on.

Complete Sets of Integers

In general, a set $A = a_1 < a_2 < a_3 < \dots$ of distinct integers will be called *complete* if there is an integer N such that every integer greater than N can be represented as a sum of distinct elements of A . For example, powers of 2 and primes are complete sets with $N = 0$ and $N = 6$, respectively.

Representing integers as a sum of some other “special” integers is an old and interesting topic in mathematics. For example, Bachet conjectured and Lagrange proved that every integer can be represented as a sum of at most 4 perfect squares. For example, $3 = 1^2 + 1^2 + 1^2$, and $31 = 5^2 + 2^2 + 1^2 + 1^2$. Of course, in Lagrange’s theorem repetitions are allowed, and, for example, 3 cannot be represented as a sum of *distinct* squares. However, Sprague [359] proved in 1948 that every integer greater than 128 is in fact a sum of distinct squares, for example, $129 = 10^2 + 5^2 + 2^2$, hence the set of perfect squares is another example of a complete set. Moreover, for any fixed k , the sequence of k -th powers $A = \{1^k, 2^k, 3^k, \dots\}$ is also a complete set!

A Set Which is Not Complete Because It is Too “Small”

What are examples of sequences which are not complete, and hence cannot be used as nominals for coin systems? One example is the sequence $A = \{1, 3, 3^2, \dots, 3^k, \dots\}$ of powers of 3. For example, the numbers 2, 5, 6, 7, 8, and so on, are not representable as a sum of distinct elements of A . To prove this, for example, for 8, we can notice that $3^k > 8$ for any $k \geq 2$, hence only 1 and 3 can be used to represent 8, but $1 + 3 < 8$, a contradiction. Similarly, for any k , numbers of the form $3^k - 1$ are never representable in this way, because coins with nominals from 3^k and larger are too big, while the sum $1 + 3 + \dots + 3^{k-1}$ of the remaining ones is too small.

An intuitive reason for the incompleteness of this set is that it contains “too few” elements. Formally, for any set A , let $A(n)$ be the number of elements in A not exceeding n . Then, for $A = \{1, 3, 3^2, \dots, 3^k, \dots\}$, $A(3^k - 1) = k$, and, in general, $A(n)$ is approximately $\log_3 n$.

“Larger” Sets Which Are Not Complete

A less trivial example is: let $m = 8$, and let A contain all integers from $m/4$ to $m/2$ (that is, 2, 3, and 4), from $m^2/4$ to $m^2/2$ (that is, from 16 to 32), from $m^4/4$ to $m^4/2$ (that is, from 1024 to 2048), from $m^8/4$ to $m^8/2$, and so on, from $m^{2^i}/4$ to $m^{2^i}/2$ for every i . Then A is not complete based on an argument similar to that for the powers of 3: integers of the form $m^{2^i}/4 - 1$, that is, 1, 15, 1023, and so on, are not representable as a sum of distinct elements of A , because elements from $m^{2^i}/4$ onwards are too large, while the sum of all other elements is too small.

For this set, and $n = m^{2^i}/4 - 1$, $A(n) \geq m^{2^{i-1}}/4 > \sqrt{n}/2$, and one can check that the inequality $A(n) > \sqrt{n}/2$ remains correct for any other $n \geq 2$. For large n , $\sqrt{n}/2$ is much bigger than $\log_3 n$, so this set contains much more elements than in the previous example, but it is still not complete.

Can there be incomplete sets which contain even more elements? Of course, and the trivial example is the set A of all even numbers. This set contains every second integer, $A(n) \approx n/2$, which is much more than $\sqrt{n}/2$ (for large n), but this set is obviously incomplete because no odd integer can be represented as a sum of even integers. A similar example is the set $A = \{1, 3, 6, 9, 12, \dots\}$ of multiples of 3 together with 1: no number of the form $3k + 2$ can be represented as a sum of distinct elements of such A .

There Are No Other Reasons for Being Not Complete!

Sets of the form $a + bk$, with a, b fixed and $k = 0, 1, 2, \dots$ are called arithmetic progressions. The examples above indicate two reasons why a set may not be complete: first, it can contain “too few” elements, and second, it may exclude some arithmetic progression for “obvious” divisibility reasons. The following theorem of Szemerédi and Vu [366] proves that these are “the only reasons”: any set which

contains “enough” elements and does not “avoid” arithmetic progression must be complete!

Theorem 6.1 *There is a positive constant c such that the following holds. Any increasing sequence $A = \{a_1 < a_2 < a_3 < \dots\}$ satisfying*

- (a) $A(n) \geq c\sqrt{n}$, and
- (b) *for any integers $a, b > 0$ there is an integer k such that $a + bk$ is representable as a sum of distinct elements of A ,*

is complete.

Theorem 6.1 confirms a conjecture made by Erdős [133] in 1962, and is very useful and universal. To establish the completeness of different special sets, such as primes, one can just verify conditions (a) and (b), instead of searching for a different proof in each case. The example above demonstrates the existence of a non-complete set with $A(n) > \sqrt{n}/2$, hence the bound in (a) is optimal up to the constant factor, and condition (b) is also unavoidable. This implies that Theorem 6.1 is the best result in this direction one can hope for.

Reference

E. Szemerédi and V. Vu, Finite and infinite arithmetic progressions in sumsets, *Annals of Mathematics* **163**-1, (2006), 1–35.

6.2 Counting Number Fields of Bounded Discriminant

Fields: Definition and Basic Examples

Addition and multiplication of real numbers satisfy a number of natural properties.

- (a) Associativity: $a + (b + c) = (a + b) + c$ and $a \cdot (b \cdot c) = (a \cdot b) \cdot c$;
- (b) Commutativity: $a + b = b + a$ and $a \cdot b = b \cdot a$;
- (c) Identities: There exist two different numbers, 0 and 1, such that $a + 0 = a$ and $a \cdot 1 = a$;
- (d) Additive inverse: for every a , there exists a number, denoted $-a$, such that $a + (-a) = 0$;
- (e) Multiplicative inverse: for every $a \neq 0$, there exists a number, denoted $1/a$, such that $a \cdot (1/a) = 1$;
- (f) Distributivity: $a \cdot (b + c) = a \cdot b + a \cdot c$.

Any set F with two operations $+$ and \cdot on it, satisfying the properties (a)–(f), is called a *field*. Hence, the set \mathbb{R} of all real numbers is an example of a field. Another example is the set \mathbb{Q} of all rational numbers, with the usual addition and

multiplication. On the other hand, the set \mathbb{Z} of all integers is not a field, property (e) fails: for example, for the integer $a = 2$ there is no integer b such that $a \cdot b = 1$.

Some Smallest Possible Fields

Are there examples of fields which are even smaller than \mathbb{Q} ? Sure! The smallest field is in fact a two-element set $F = \{0, 1\}$, with operations defined as $0 + 0 = 0, 0 + 1 = 1 + 0 = 1, 1 + 1 = 0, 0 \cdot 0 = 0 \cdot 1 = 1 \cdot 0 = 0, 1 \cdot 1 = 1$. One can check directly that properties (a)–(f) hold with $-0 = 0, -1 = 1$, and $1/1 = 1$. A possible interpretation of this field is that 0 denotes all even integers, and 1 denotes all odd integers. Then, for example, $1 \cdot 0 = 0$ corresponds to the fact that the product of an odd and even integer is even, while $1 + 1 = 0$ reads as “the sum of two odd integers is even”.

However, $1 + 1 = 0$ does not hold with our “regular” addition. Are there subsets of real numbers, other than \mathbb{Q} , which form a field with the *usual* addition and multiplication? Well, any such field F should contain 0 and 1 by (c), hence it should also contain $1 + 1 = 2, 2 + 1 = 3$, and so on. Then, it should contain negative integers $-1, -2, \dots$ by (d), fractions in the form $1/m$ (for integers $m \neq 0$) by (e), thus also numbers $n/m = n \cdot (1/m)$, $n, m \in \mathbb{Z}, m \neq 0$, that is, all rational numbers! Hence, for fields with the usual operations, the field \mathbb{Q} of rational numbers is the smallest one.

The Smallest Field Containing $\sqrt{2}$

One may ask what is the smallest field F with usual operations containing some irrational number, say, $\sqrt{2}$. By (c), F should contain 0 and 1, and hence, by (d) and (e), all rational numbers as well. Next, we also need to include all numbers of the form $a + b\sqrt{2}, a, b \in \mathbb{Q}$, to be able to perform addition and multiplication with $\sqrt{2}$. It turns out that this set $F = \{a + b\sqrt{2}, a, b \in \mathbb{Q}\}$ is indeed a field! For any $u = a + b\sqrt{2}$ and $v = c + d\sqrt{2}$, F contains their sum $u + v = (a + c) + (b + d)\sqrt{2}$ and product $u \cdot v = (a + b\sqrt{2})(c + d\sqrt{2}) = (ac + 2bd) + (ad + bc)\sqrt{2}$, and properties (a), (b), and (f) hold. Next, F includes 0 and 1, and also the inverse $-u = (-a) + (-b)\sqrt{2}$ for every $u = a + b\sqrt{2}$, hence (c) and (d) follow. But what about the multiplicative inverse in (e)? Can, for example, $1/(2 + 3\sqrt{2})$ be written as $a + b\sqrt{2}$ for rational a, b ? It turns out, it can. We can multiply the numerator and denominator by $(2 - 3\sqrt{2})$, and, using the fact that $(2 - 3\sqrt{2})(2 + 3\sqrt{2}) = 2^2 - (3\sqrt{2})^2 = -14$, write $1/(2 + 3\sqrt{2})$ as $(2 - 3\sqrt{2})/(-14) = (-1/7) + (3/14)\sqrt{2}$. The same trick works in general, and (e) follows.

Number Fields

$\sqrt{2}$ is a solution of the quadratic equation $x^2 - 2 = 0$. Let us form the smallest field F with the usual operations containing a root α of some cubic equation, say $x^3 + x + 1 = 0$. As before, F should contain all rational numbers, and, more generally, all numbers of the form $a + b\alpha$, $a, b \in \mathbb{Q}$. However, we should also be able to multiply α by α and include α^2 , and therefore F should contain all numbers of the form $a + b\alpha + c\alpha^2$, $a, b, c \in \mathbb{Q}$. This is indeed a field, no more elements are needed. For example, $\alpha \cdot \alpha^2 = \alpha^3$, but we do not need to include α^3 separately, because $\alpha^3 + \alpha + 1 = 0$, hence $\alpha^3 = -1 - \alpha$ is already included. Similarly, $\alpha^2 \cdot \alpha^2 = \alpha^4$, but $\alpha(\alpha^3 + \alpha + 1) = 0$ implies that $\alpha^4 = -\alpha - \alpha^2 \in F$. Moreover, $\alpha^3 + \alpha + 1 = 0$ implies that the inverse $1/\alpha = -1 - \alpha^2$ is also an element of F , and one can show that the inverse $1/z$ of any $z = a + b\alpha + c\alpha^2$ belongs to F by a similar argument.

Similarly, if α is a root of any equation of degree n with rational coefficients,¹ the smallest field with the usual operations containing it is $F = F(\alpha) = \{a_0 + a_1\alpha + \dots + a_{n-1}\alpha^{n-1}, a_0, a_1, \dots, a_{n-1} \in \mathbb{Q}\}$. Such fields are called *number fields* of degree n . For example, $F(\sqrt{2}) = \{a + b\sqrt{2}, a, b \in \mathbb{Q}\}$ is an example of a number field of degree 2.

How Many Different Number Fields Are There?

What if we consider the equation $x^2 - 8 = 0$, so that $\alpha = \sqrt{8}$? Then $F(\sqrt{8})$ is the set of all numbers representable as $z = a + b\sqrt{8}$ for rational a, b . However, $\sqrt{8} = 2\sqrt{2}$, and $z = a + b\sqrt{8} = a + 2b\sqrt{2} \in F(\sqrt{2})$. Conversely, any element $z = a + b\sqrt{2}$ of $F(\sqrt{2})$ can be written as $a + (b/2)\sqrt{8}$. Hence, in fact, $F(\sqrt{2})$ and $F(\sqrt{8})$ are the same number field, just written down in two different ways! In this case, we say that fields $F(\sqrt{2})$ and $F(\sqrt{8})$ are *isomorphic*.

A basic question in this theory is how many *different*, that is, non-isomorphic number fields one can form. Stated in this form, the question is trivial, and the answer is: infinitely many. For example, for any prime p , one can form a number field $F(\sqrt{p}) = \{a + b\sqrt{p}, a, b \in \mathbb{Q}\}$, and all these number fields are different.

A similar situation may arise if we just ask “how many primes exist?”, or “how many powers of 2 exist?”. The trivial answer is “infinitely many” to both questions. However, the correct formulation of this question is: given a large integer N , how many primes (powers of 2) exist with absolute value at most N ? Then the question makes sense, and the prime number theorem states that there are about $N / \ln N$ primes. In contrast, there are just about $\log_2 N$ powers of two.

¹Of course, the root $\alpha = \sqrt{2}$ of $x^2 - 2 = 0$ is also the root of, say, $x^{10} - 2x^8 = 0$. For concreteness, we will assume that the degree n is chosen to be the smallest possible.

Number Fields with Fixed Degree and Bounded Discriminant

Number fields do not have an “absolute value”. However, one can associate to each number field K an integer d_K , which is called the *discriminant* of K . See Sect. 5.9 for a detailed discussion and the exact definition of the discriminant. For our purpose, what is important is that, for all integers n and X , the number $N_n(X)$ of different number fields with degree n and absolute value of the discriminant at most X is finite. Hence, our basic question can now be rigorously formulated: how large is $N_n(X)$? Theorem 5.9 discussed in Sect. 5.9 answers this question for $n = 4$, but for general n the question is very difficult and far from being understood. There is a conjecture that, for every fixed n , $N_n(X)$ should grow as a linear function of X , but the best general upper bound known before 2006 was

$$N_n(X) \leq C_n X^{(n+2)/4},$$

where C_n is a constant which depends on n , see [335]. The following theorem of J. Ellenberg and A. Venkatesh [126] provides a much better bound.

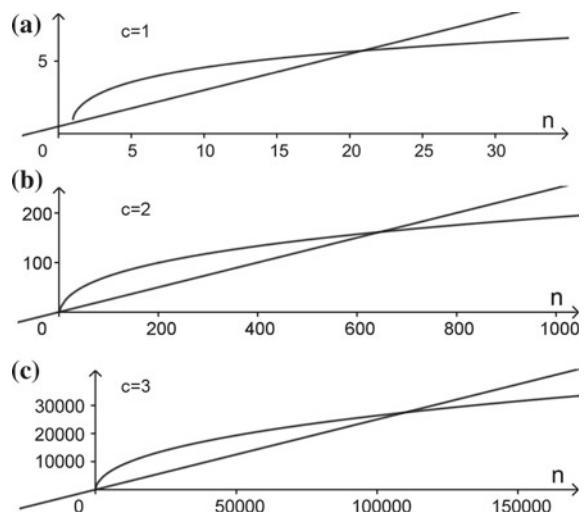
Theorem 6.2 *For all $n > 2$, we have*

$$N_n(X) \leq B_n X^{\exp(C\sqrt{\ln n})},$$

where B_n is a constant depending only on n , and C is an absolute constant.

For example, for $n = 1,000$ the previous bound gave the estimate $N_n(X) \leq C_n X^{250.5}$, while Theorem 6.2 with $C = 1$ would imply $N_n(X) \leq C_n X^{13.85}$. Of course, one cannot assume that $C = 1$, but, for any C , the exponent $\exp(C\sqrt{\ln n})$ in Theorem 6.2 grows with n much slower than $(n + 2)/4$, see Fig. 6.2.

Fig. 6.2 Rate of growth of $\exp(C\sqrt{\ln n})$ versus $(n + 2)/4$ for different values of C



Reference

J. Ellenberg and A. Venkatesh, The number of extensions of a number field with fixed degree and bounded discriminant, *Annals of Mathematics* **163**-2, (2006), 723–741.

6.3 Perfect Powers in Fibonacci and Lucas Sequences

The Sequence of Fibonacci Numbers

Perhaps the most famous sequence of integers in mathematics is the sequence of *Fibonacci numbers*. It is defined by the following rules: $F_0 = 0$, $F_1 = 1$,

$$F_{n+2} = F_{n+1} + F_n, \quad \text{for } n = 0, 1, 2, 3, \dots \quad (6.1)$$

In particular, $F_2 = F_1 + F_0 = 1 + 0 = 1$, $F_3 = F_2 + F_1 = 1 + 1 = 2$, $F_4 = 2 + 1 = 3$, $F_5 = 3 + 2 = 5$, and so on, so the first terms are

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, \dots$$

The numbers are named after the Italian mathematician who wrote about them in his 1202(!) book, although the sequence was known before then. Fibonacci numbers are so important and popular in mathematics that there exists a special journal, named the *Fibonacci Quarterly*, in which all publications are devoted to the study of Fibonacci numbers only!

An Explicit Formula for F_n

For example, one of the nice results about Fibonacci numbers is that there exists an explicit formula for F_n . Numerical experiments suggest that F_n grows exponentially fast, so let us look for a formula in the form $F_n = C\phi^n$ for some constants C, ϕ . Substituting this into (6.1) results in $C\phi^{n+2} = C\phi^{n+1} + C\phi^n$, or $\phi^2 = \phi + 1$, which implies that $\phi = \frac{1}{2}(1 + \sqrt{5})$ or $\frac{1}{2}(1 - \sqrt{5})$. However, the formula $F_n = C\phi^n$ does not satisfy the requirements $F_0 = 0$, $F_1 = 1$. Indeed, $F_0 = 0$ would imply that $C\phi^0 = 0$, or $C = 0$, but then $F_1 = C\phi^1 = 0 \neq 1$, a contradiction.

Let us try a slightly more complicated formula, $F_n = A\phi^n + B\psi^n$. Again, substituting this into (6.1) results in $A(\phi^2 - \phi - 1) + B(\psi^2 - \psi - 1) = 0$, hence we can select $\phi = \frac{1}{2}(1 + \sqrt{5})$ and $\psi = \frac{1}{2}(1 - \sqrt{5})$. Now, $F_0 = 0$ implies that $A\phi^0 + B\psi^0 = 0$, or $B = -A$. Finally, $F_1 = 1$ implies that $A\phi^1 + B\psi^1 = 1$, or $A\phi + (-A)\psi = 1$, hence $A = \frac{1}{\phi - \psi} = \frac{1}{\sqrt{5}}$. Hence, we have derived the formula

$$F_n = \frac{\phi^n - \psi^n}{\sqrt{5}}, \quad \text{where } \phi = \frac{1}{2}(1 + \sqrt{5}), \quad \psi = \frac{1}{2}(1 - \sqrt{5}).$$

It may be difficult to believe that a nice integer sequence has a formula in terms of irrational numbers ϕ and ψ , but this is indeed the case. All irrationals “magically” cancel out, resulting in an integer. For example,

$$F_2 = \frac{1}{\sqrt{5}} \left[\left(\frac{1 + \sqrt{5}}{2} \right)^2 - \left(\frac{1 - \sqrt{5}}{2} \right)^2 \right] = \frac{1}{\sqrt{5}} \cdot \frac{4\sqrt{5}}{2^2} = 1, \quad (6.2)$$

and so on. In fact, $\psi = \frac{1}{2}(1 - \sqrt{5}) \approx -0.6$, hence ϕ^n is very small for large n , and (6.2) “simplifies” to

$$F_n \approx \frac{\phi^n}{\sqrt{5}}.$$

In fact, F_n is just the nearest integer to $\frac{\phi^n}{\sqrt{5}}$. For example, $\frac{\phi^{10}}{\sqrt{5}} \approx 55.0036$ and $F_{10} = 55$.

Perfect Powers in the Fibonacci Sequence

While the derivation of formula (6.2) is simple, there are some basic questions about Fibonacci numbers which are highly non-trivial. For example, out of the first 13 Fibonacci numbers there are five which are perfect powers of an integer, that is, representable in the form k^p for some integers k and $p \geq 2$. These are $F_0 = 0 = 0^p$, $F_1 = F_2 = 1 = 1^p$, $F_6 = 8 = 2^3$, and $F_{12} = 144 = 12^2$, see Fig. 6.3a. One can use a computer to compute millions of further Fibonacci numbers, but there will be no other perfect powers among them. Based on this, people conjectured that there are no more. In 1951, Ljunggren [248] proved that there are no more perfect squares in the sequence. In 1969, London and Finkelstein [249] proved that there are no more perfect cubes. Continuing this line of research, people proved that there are no more perfect powers with $p \leq 17$. Interestingly, in 2001, Pethő [303] showed that there are no more perfect powers with $p \geq 5.1 \cdot 10^{17}$, hence, by 2006, the problem was solved for very small values of p and for very large ones, but not for the general case!

Perfect Powers in the Lucas Sequence

Of course, $F_0 = 0$ and $F_1 = 1$ are not the only initial two terms you can start with. For example, in the 19th century the mathematician François Lucas suggested to start with $L_0 = 2$ and $L_1 = 1$, and then apply the same rule (6.1), resulting in the sequence

$$2, 1, 3, 4, 7, 11, 18, 29, 47, 76, \dots$$

Once again, a general formula can easily be derived, which in this case is

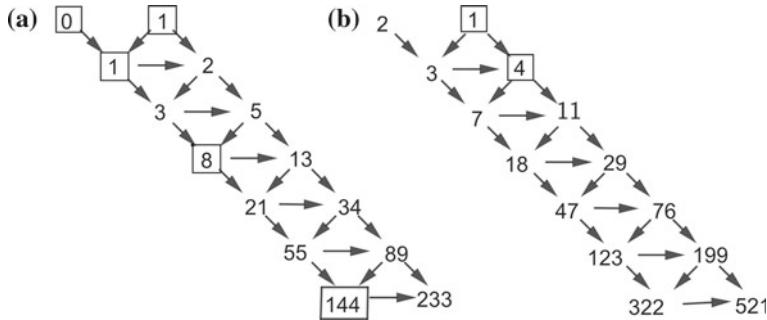


Fig. 6.3 Perfect powers in the Fibonacci and Lucas sequences

$$L_n = \phi^n + (1 - \phi)^n, \quad \phi = \frac{1}{2}(1 + \sqrt{5}),$$

but, similar to Fibonacci numbers, it is much more difficult to find all the perfect powers in this sequence. We can easily see that $L_1 = 1 = 1^p$ and $L_3 = 4 = 2^2$, see Fig. 6.3b, but are there others? By 2006, it was known that there are no more perfect squares, and no more cubes, and also no more perfect powers with $p \geq 13222$. So, once again, the problem was solved for very small p , and for large p , but not for “intermediate” values of p .

The Solution to Both Questions

In 2006, both questions were answered in full [81]:

- Theorem 6.3** (a) *The only perfect powers in the Fibonacci sequence are $F_0 = 0$, $F_1 = 1$, $F_2 = 1$, $F_6 = 8$, and $F_{12} = 144$.*
 (b) *The only perfect powers in the Lucas sequence are $L_1 = 1$ and $L_3 = 4$.*

As usual, to prove a great theorem, the authors needed to develop techniques which can be applied to similar problems. For example, consider the sequence $G_n = n^2 + 7$, that is, $G_0 = 0^2 + 7 = 7$, $G_1 = 1^2 + 7 = 8$, $G_2 = 2^2 + 7 = 11$, and so on. The first terms are

$$7, 8, 11, 16, 23, 32, 43, 56, 71, 90, 107, 128, 151, \dots$$

We can see several perfect powers, namely $G_1 = 8 = 2^3$, $G_3 = 3^2 + 7 = 16 = 2^4$, $G_5 = 5^2 + 7 = 32 = 2^5$, $G_{11} = 11^2 + 7 = 128 = 2^7$. Are there more? In other words, what are the integer solutions to the equation $x^2 + 7 = y^p$, $p \geq 2$? By 2006, this was not known. Using a similar technique to the one in the proof of Theorem 6.3, the same authors [82] proved that the only other solution is $181^2 + 7 = 32768 = 2^{15}$. Moreover, they solved a similar equation $x^2 + D = y^p$, $p \geq 3$, which they called the “Lebesgue–Nagell equation”, for all values of D between 1 and 100. For example, for $D = 100$ the integer solutions are $5^2 + 100 = 5^3$, $30^2 + 100 = 10^3$,

$198^2 + 100 = 34^3$, and $55^2 + 100 = 5^5$. Of course, the corresponding negative solutions $x = -5, -30, -198$ and -55 also work, but there are no more.

Reference

Y. Bugeaud, M. Mignotte, S. Siksek, Classical and modular approaches to exponential Diophantine equations I. Fibonacci and Lucas perfect powers, *Annals of Mathematics* **163**-3, (2006), 969–1018.

6.4 A Characterization of Perfect Graphs

Using Graph Colouring for Party Organization

Imagine, like in Sect. 5.11, that you are organising a party, but some of your guests do not like each other, and would like to sit at different tables. The question is what is the minimal number of tables you would need. For example, with 5 guests, such that Anna dislikes Bob, Bob dislikes Claire, Claire—David, David—Eva, and Eva dislikes Anna, 3 tables surely suffice: for example, Anna could sit with Claire, Bob with David, and Eva alone. However, 2 tables is not sufficient: if Anna sits at the first one, then Bob needs to be near the second one, but then Claire near the first, David—second, Eva—first, but she dislikes Anna...

An appropriate mathematical language for this problem is that of *graph theory*: you can represent each guest as a point on the plane (called a *vertex*), and connect two points by a line (called an *edge*) if the corresponding guests do not like each other, see Fig. 6.4. Then, distributing guests into k tables is equivalent to colouring the corresponding points/vertices into k colours, and the requirement that “guests which do not like each other should sit near different tables” translates into “every two vertices connected by an edge must have a different colour”—such a colouring is called *proper*. Our question now is: what is the minimal number of colours we need for a proper colouring?

The set of vertices, some of which are connected by an edge, is called a *graph*, and the minimal number of colours we are looking for is called the *chromatic number* of the graph. In the example above with 5 guests, they can be represented as vertices A, B, C, D, E , with edges AB, BC, CD, DE , and EA . This graph can be drawn on the plane as a pentagon. The task then is to colour the vertices of the pentagon so that no two adjacent vertices have the same colour. The answer is that the minimal number of colours—the chromatic number of the pentagon—is 3, see Fig. 6.4c.

Other Applications of Graph Colouring

Of course, graph colouring has many more applications than “just” party organization. For example, assume that we need to schedule university lectures for different subjects, e.g. Algebra, Business Management, Computer Science, Data Analysis,

Economics, etc. If there are students registered for, say, Algebra and Business Management, then these two lectures cannot be scheduled at the same time. The question is what is the minimal number of time slots we need. Once again, the subjects can be represented as graph vertices, and two vertices are connected by edges if and only if the two classes cannot occupy the same time slot. Then any “legitimate” lecture scheduling corresponds to a proper colouring of the resulting graph, and we are faced with *exactly* the same problem of finding a proper colouring with the minimal number of colours. Hence, if we could learn how to allocate tables to guests, we would know how to colour graphs, and could therefore solve various other problems, arising from completely different applications. This is the beauty and power of mathematics.

How to Convince Yourself That Your Solution is the Best Possible?

In general, allocating guests/colouring graphs in an optimal way is not easy, and it is also not easy to *prove* that the colouring you found is optimal. Imagine, for example, that you have found a way to allocate your guests into 4 tables. But how could you convince the restaurant (and yourself!) that you really need all 4 tables, and there is no smart way to allocate all guests in a conflict-free way using just 3 tables? Well, sometimes this is easy. For example, if you have 4 guests who all dislike each other, then they all need to sit at different tables, hence clearly 3 tables would not suffice. More generally, if you can name k guests who are all enemies of each other, then this would be a convincing proof that we need at least k tables.

In general, the size of the largest set of vertices which are all connected to each other is called the *clique number* of a graph. Because all these vertices should have a different colour, the chromatic number is always greater than or equal to the clique number. If equality holds, we can be absolutely sure that our colouring is optimal, and can easily prove this to everyone.

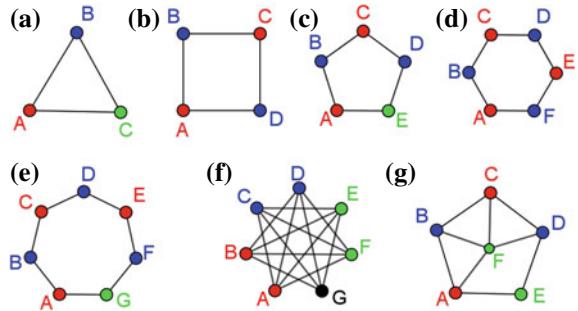
Colouring Graphs in the Form of n -gons

For example, a graph with vertices A, B, C and edges AB, BC , and CA , which can be drawn in the plane as a triangle, requires 3 colours, see Fig. 6.4a, and this is obviously the best possible because its clique number is also 3. For a graph with vertices A, B, C, D and edges AB, BC, CD , and DA (a quadrangle), we can colour A and C in red, while B and D in blue, see Fig. 6.4b, hence its chromatic number is 2, and its clique number is 2 as well.

However, this convenient equality (chromatic number = clique number) does not always hold. In particular, the pentagon in Fig. 6.4c contains no triangle, hence its clique number is 2. However, as discussed above, for a pentagon two colours do not suffice, and its chromatic number is equal to 3.

For a hexagon $ABCDEF$, we can colour A, C, E red, and B, D, F blue, see Fig. 6.4d, hence chromatic number = clique number = 2. However, for a heptagon $ABCDEFG$, colouring A red, B blue, C red, and so on, would result in G red, the

Fig. 6.4 Some examples of graph colourings



same as A , a contradiction, hence the chromatic number is 3, see Fig. 6.4e, while the clique number is 2. The same problem occurs with all graphs representable as an n -gon with odd $n \geq 5$. Such graphs are called *odd holes*.

A Delicate Analysis of Odd Antiholes

As another example, consider a graph with 7 vertices representable as a heptagon $ABCDEFG$ with all diagonals drawn, but all sides erased. So, A is connected to C, D, E, F , but not to B and G , and so on, see Fig. 6.4f. Because vertices A, C , and E are all connected to each other, it is trivial that at least 3 colours are needed. However, it turns out that 3 is not enough and we really need 4. Why? Well, let A be red, C blue, then E is connected to A and C and should therefore have a third colour, say green, and F should be green as well for the same reason. Then D (connected to red A and green F) should be blue, B (connected to blue D and green F) should be red, but then G is connected to red B , blue C (and D), and green E , so we should use a fourth colour for it.

Complicated? Yes. And the reason is that, for this graph, the chromatic number is 4, but the clique number is only 3: there are no 4 vertices which are all connected to each other, hence there is no easy demonstration that 3 colours are not enough, and a delicate case-by-case analysis is required. The same situation happens with any graph formed from the diagonals of an n -gon with any odd $n \geq 5$. Such graphs are called *odd antiholes*.

A Problem with Non-arriving Guests and Perfect Graphs

Our final example is a party with 5 guests A, B, C, D, E forming a pentagon, and one “special guest” F who hates A, B, C, D (but not E). Then we can colour A, C red, B, D blue, E, F green, hence 3 colours suffices, see Fig. 6.4g. Because there are triples (for example, A, B, F) of pairwise haters, we can easily convince the restaurant workers that this is optimal, 2 tables would not be enough. However, imagine now that the “special guest” F did not arrive! Then, for the remaining 5 guests forming

a pentagon, we would still need 3 tables, but there is no easy demonstration that 2 tables do not suffice.

This motivates the following definition. A graph G is called *perfect* if its chromatic number is equal to its clique number, and, moreover, this is the case for *any induced subgraph* of it. Here, an induced subgraph is a graph resulting from deleting some vertices of G together with the corresponding edges. If the graph corresponding to our party is perfect, then we can always easily prove that our optimal k -colouring is indeed optimal, just by demonstrating a group of k guests who all dislike each other, and we can do this even if some guests do not arrive. Moreover, there is an algorithm [252] (discovered in 1988) which can efficiently *find* an optimal colouring for any perfect graph!

The Proof of Berge's Conjecture

However, what are the perfect graphs, what do they look like? For a pentagon, the chromatic number is 3, while the clique number is 2, hence any graph containing a pentagon as an induced subgraph is not perfect by definition. For the same reason any graph containing an odd hole or odd antihole as an induced subgraph cannot be perfect. In 1961, Berge [44] conjectured the amazing fact that *all* other graphs are in fact perfect! The following theorem of Chudnovsky et al. [92] confirms this conjecture!

Theorem 6.4 *A graph G is perfect if and only if it does not contain an odd hole or an odd antihole as an induced subgraph.*

Hence, if you need to colour a graph, arising from a party organization, or university scheduling, or from any other application, you could check if it contains an odd hole or odd antihole as an induced subgraph (in 2005 Chudnovsky et al. [91] developed an algorithm to do this efficiently), and if not, you can be sure that your graph is perfect.

Reference

M. Chudnovsky, N. Robertson, P. Seymour, R. Thomas, The strong perfect graph theorem, *Annals of Mathematics* **164**-1, (2006), 51–229.

6.5 Littlewood's Conjecture Holds Outside of a Set of Dimension 0

Rational Approximation of Irrational Numbers

The area of a square $ABCD$ with side length $|AB| = 3$ can be calculated and written down exactly: it is $3^2 = 9$. Similarly, the area of the triangle ABC is $\frac{9}{2} = 4.5$. It is not an integer, but can be written down exactly either in finite decimal form (4.5),

or as a ratio of two integers ($\frac{a}{b}$). Numbers which can be written as a ratio $\frac{a}{b}$ of two integers a and b are called *rational* numbers.

However, the area of a circle with radius 1, known as the mathematical constant π , is not a rational number. It has an *infinite* decimal expansion starting from 3.14159265..., and the more digits in this expansion we write down, the better approximation to π we get. However, it is impossible to write π in this way *exactly*: we would need an infinite number of digits for this! Similarly, π cannot be written down as a ratio $\frac{a}{b}$ with integers a, b . Of course, we can express it approximately, for example, $\frac{31}{10}$, or $\frac{314}{100} = \frac{157}{50}$, or $\frac{314159}{100000}$, etc. Furthermore, we can make the approximation error as small as we like. Indeed, for any denominator b , let a be the largest integer such that $\frac{a}{b} < \pi$. Then $\pi < \frac{a+1}{b}$, and therefore either $|\pi - \frac{a}{b}| < \frac{1}{2b}$, or $|\frac{a+1}{b} - \pi| < \frac{1}{2b}$. In other words, with denominator b , we can guarantee the approximation error to be less than $\frac{1}{2b}$. By selecting b sufficiently large, we can make $\frac{1}{2b}$ as small as we want. For example, to guarantee the error to be less than $\frac{1}{1000}$, we can select $b = 500$, and write $\pi \approx \frac{1571}{500} = 3.142$.

Better Approximation with “Special” Denominators

In fact, it is interesting and useful to have a good approximation to π with *small* b . For example, the approximation $\pi \approx \frac{22}{7} = 3.142857\dots$ is a little less precise than $\frac{1571}{500}$, but it surely “looks nicer”. It is exciting that we can achieve the approximation error $\frac{22}{7} - \pi \approx 0.00126 < \frac{1}{2395}$ with “just” $b = 7$ instead of $b = 395$. An even more exciting example is the approximation $\pi \approx \frac{355}{113} \approx 3.1415929\dots$, which gives the first *seven* digits of π correctly. The approximation error is just 0.00000026..., and, in fact, $\frac{355}{113}$ approximates π much better than $\frac{314159}{100000}$.

So, there are some “special” denominators b such that we can approximate π as $\frac{a}{b}$ with accuracy much-much better than the upper bound $\frac{1}{2b}$. This is not a coincidence: in the 19th century, Dirichlet proved that, for any irrational number α , there are infinitely many pairs of integers a and b such that

$$\left| \alpha - \frac{a}{b} \right| < \frac{1}{b^2}.$$

Of course, for large b , the error $\frac{1}{b^2}$ is much lower than $\frac{1}{2b}$. In 1903, Borel proved a better bound $|\alpha - \frac{a}{b}| < \frac{1}{\sqrt{5}b^2}$, but further improvements are impossible: there is a number $\phi = \frac{1+\sqrt{5}}{2}$, called the *golden ratio*, which cannot be approximated better.

Simultaneous Approximation and Littlewood's Conjecture

We have learned that, for every irrational number α , there are some special denominators b such that we can write $\alpha \approx \frac{a}{b}$ with an unusually small error. For example, for

π , denominators $b = 7$ and $b = 113$ work very well. Of course, the choice of b is not universal, and depends on the particular number we would like to approximate. For example, for $\alpha = \sqrt{5} \approx 2.236$ the best approximation with $b = 7$ is $\frac{16}{7} \approx 2.2857\dots$, with error about 0.05, which is not exciting at all: even the approximation $\frac{9}{4} = 2.25$ works better. However, maybe the denominator 113 would work for $\sqrt{5}$? After all, by Dirichlet's theorem, there are infinitely many denominators b which work well for π , and there are infinitely many b 's which work well for $\sqrt{5}$: can we find at least some b 's which work well *simultaneously* for π and $\sqrt{5}$? That is, can we write $\pi \approx \frac{a}{b}$ and $\sqrt{5} \approx \frac{c}{b}$ for some integers a, b, c , such that both approximations are “good”?

What exactly do we mean by “good” here? Well, by Dirichlet's theorem, there are infinitely many b 's such that $|\pi - \frac{a}{b}| < \frac{1}{b^2}$ for some a . Next, for every b , we can write $|\sqrt{5} - \frac{c}{b}| < \frac{1}{2b}$ for some c . This implies

$$\left| \pi - \frac{a}{b} \right| \cdot \left| \sqrt{5} - \frac{c}{b} \right| < \frac{0.5}{b^3}.$$

By using Borel's bound instead of Dirichlet's, we can prove the same inequality with constant $0.5/\sqrt{5} \approx 0.22$ instead of 0.5. Can we improve this constant further? It is very natural to conjecture that we can. In fact, the estimate $\frac{1}{2b}$ is a very rough bound, the worst-case scenario. There is absolutely no reason to believe that all denominators b which approximate π well, would, for some magical reason, approximate $\sqrt{5}$ in the worst possible way.

Motivated by informal arguments like this, in the 1930s Littlewood made the following conjecture: for every pair of real numbers α and β , and any $\varepsilon > 0$, there exist integers a, b, c such that

$$\left| \alpha - \frac{a}{b} \right| \cdot \left| \beta - \frac{c}{b} \right| < \frac{\varepsilon}{b^3}.$$

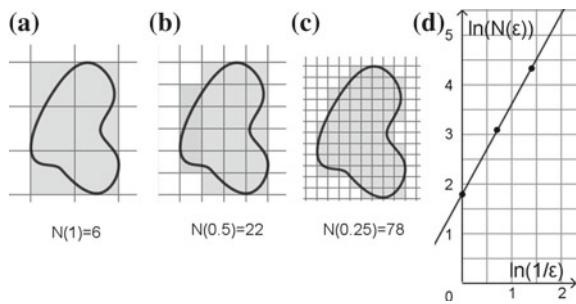
The Box Dimension for an Arbitrary Set

This conjecture turns out to be extremely difficult. So, instead of proving it for *all* pairs α and β , people tried to prove it at least for “almost all” pairs. What does this mean? Well, assume, for simplicity, that $0 \leq \alpha, \beta \leq 1$, so that all possible pairs (α, β) forms a unit square in the coordinate plane. What does it mean to say that the conjecture holds for “most”, or “almost all” points of this square?

Intuitively, there are, for example, “more” points inside the square than on its boundary. Why do we think so? Well, the interior of the square is 2-dimensional, while the boundary consists of just 4 one-dimensional lines. So, one intuitive way to judge “how big” a set is is to look at its dimension.

It is a bit tricky to formally define dimension for arbitrary sets, but it is possible to do this. The intuitive idea is that, to cover the one-dimensional interval $[0, 1]$ by small intervals of length ε , we need about $N(\varepsilon) = \frac{1}{\varepsilon}$ such intervals; to cover the two-

Fig. 6.5 Covering an area by smaller and smaller squares



dimensional unit square by small squares of side length ε , we need about $N(\varepsilon) = \left(\frac{1}{\varepsilon}\right)^2$ squares; to cover the 3-dimensional unit cube, we need $N(\varepsilon) = \left(\frac{1}{\varepsilon}\right)^3$ small cubes, etc. You will notice that the dimension is just the exponent in this expression. That is, the dimension d is the number such that $N(\varepsilon) = \left(\frac{1}{\varepsilon}\right)^d$. To find d from this equation, we need to take logarithm of both sides, and we find that $d = \ln N(\varepsilon) / \ln(1/\varepsilon)$. Because this intuition works only for small ε , we need to take the limit $\varepsilon \rightarrow 0$, and, for any set S in (for example) the plane, define

$$\dim(S) := \lim_{\varepsilon \rightarrow 0} \frac{\ln N(\varepsilon)}{\ln(1/\varepsilon)},$$

where $N(\varepsilon)$ is the number of squares with side length ε needed to cover S , see Fig. 6.5. The number $\dim(S)$ is called the *box dimension* of the set S .

The Set of Exceptions to Littlewood's Conjecture

Of course, one may ask why we have covered S by squares, but not by, say, small circles or small triangles. If we allow any shapes, then the same idea leads to the notion of *Hausdorff dimension*, see Sect. 3.7 for formal definition and related discussion.

Now we are ready to formulate the theorem of this section, proved in [125].

Theorem 6.5 *The set of all pairs (α, β) for which Littlewood's conjecture does not hold has Hausdorff dimension 0. In fact, this set can be written as a union of sets $S_1 \cup S_2 \cup S_3 \cup \dots$, where each S_i has box dimension 0.*

Of course, we would expect Littlewood's conjecture to be true, so the set of pairs (α, β) for which it does *not* hold is in fact an empty set. While Theorem 6.5 cannot prove this, it guarantees that the set of possible counterexamples is extremely small in a well-defined sense.

Reference

M. Einsiedler, A. Katok, E. Lindenstrauss, Invariant measures and the set of exceptions to Littlewood's conjecture, *Annals of Mathematics* **164**-2, (2006), 513–560.

6.6 The Connection Between Metric Entropy and Combinatorial Dimension

Several Ways to Measure “Size”

How would you compare the “size” of two regions in the plane, for example, two cities? The most obvious way is to compare their areas. However, the question is not as trivial as it seems if the shapes of the cities are unusual. For example, the Ukrainian city Kryvyi Rih extends 126 km from north to south, and it is the longest in Europe! However, its area is just about 430 square kilometres, from which we can conclude that its average width is just about $430/126 \approx 3.4$ km. Any city in the form of a square with side length 21 km would have a large area. However, if you forget about area and just use your intuition, would you really call such a city “larger” than 126-km-long Kryvyi Rih? Or, if we think about theoretical geometry, would you call the tiny square with side length 10^{-3} mm (and area 10^{-6}) “larger” than a line segment of length 10^6 km (and area 0)?

An alternative way to estimate the “size” of a city is to count how many points it contains which are all at large distance, say 10 km, from each other. In the square city with side length 21 km, we can fit 9 such points (vertices of the square, middles of the sides, and the center of the square), while in Kryvyi Rih we can easily fit at least 13 such points along the straight line from north to south, and in fact more using the fact that it also has a non-zero width.

More formally, for a subset S of the plane, and a real number $t > 0$, let $N(S, t)$ be the maximal number of points in S such that the distance between any of them is at least t . For example, if S is a square with integer side length m , then $N(S, 1) = (m + 1)^2$; if S is a straight line of integer length k , then $N(S, 1) = k + 1$. So, in this way we can compare the “sizes” of a square and a line in a more interesting way than just saying that any square, no matter how tiny, is larger than any line, no matter how long.

Food Baskets and “Sizes” of k -Dimensional Sets

Of course, we may talk not only about cities. For example, assume that we go to a shop to buy some food. For each possible food basket, we are interested in some two parameters, say, how much vitamins A and B it contains. Then a food basket containing amount a of vitamin A and amount b of vitamin B can be represented as a point (a, b) on the plane, and the set S of all possible food baskets becomes just a set of such points. Now, we can measure the “distance” between food baskets X and Y as the distance $d(X, Y)$ between the corresponding points, and then $N(S, t)$ is the maximal number of *different* (that is, at “distance” at least t from each other) food baskets we can buy.

In the last example, it is a bit strange that we can completely describe any food basket using just two parameters. In reality, there may be any number (say, k) parameters

of interest, and then each food basket can be described as a “point in k -dimensional space” with k coordinates. The distance between points $X = (x_1, x_2, \dots, x_k)$ and $Y = (y_1, y_2, \dots, y_k)$ can be measured as

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_k - y_k)^2},$$

and then, as above

$$N(S, t) = \{\max n : \text{exists } X_1, X_2, \dots, X_n \in S, \text{ such that } d(X_i, X_j) \geq t, i \neq j\}. \quad (6.3)$$

How “Big” is a Set of Functions?

In other applications, it is sometimes important to estimate how “big” the set of functions from some set Ω to real line is. For example, let us consider functions from $\Omega = [0, 1]$ to the real line, and let S_1 be the set of linear functions $f(x) = a + bx$, $-1 \leq a, b \leq 1$, while S_2 is the set functions $f(x) = a + bx^2$, $-1 \leq a, b \leq 1$. Which one is bigger? Sets of functions, like S_1 and S_2 , do not have a volume, so traditional methods are not applicable. However, we still can use the above idea and count how many “different” functions there are in each set. To start, we can select any specific point $x^* \in [0, 1]$, say $x^* = 0.5$, and say that functions f and g are “different” if $|f(x^*) - g(x^*)| \geq t$. With this definition and $t = 1$, S_1 contains 4 functions ($1+x$, $1-x$, $-1+x$, $-1-x$) at “distance” at least 1 from each other, while S_2 has at most 3 (for example, $1+x^2$, 0 , $-1-x^2$), see Fig. 6.6a, b. But S_2 could be “large” if we select a different x^* . More importantly, it is inadequate to compare functions by using their values at one point only.

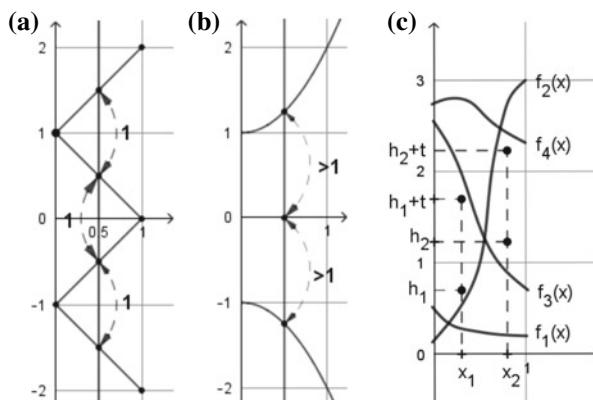


Fig. 6.6 Functions which are “significantly different” in various senses

The Metric Entropy of a Set of Functions

To improve the method, let us select some k points $x_1, x_2, \dots, x_k \in \Omega$ (possibly with repetitions), and, for any two functions f and g , define

$$d_{x_1, \dots, x_k}(f, g) = \frac{1}{\sqrt{k}} \sqrt{(f(x_1) - g(x_1))^2 + (f(x_2) - g(x_2))^2 + \dots + (f(x_k) - g(x_k))^2},$$

where $\frac{1}{\sqrt{k}}$ is a scaling factor to ensure that $d_{x_1, \dots, x_1}(f, g)$, where the point x_1 is repeated k times, is the same as $d_{x_1}(f, g)$. In words, $d_{x_1, \dots, x_k}(f, g)$ measures how different functions f and g are at points x_1, \dots, x_k . Then, for any set S of functions, and $t > 0$, let $N_{x_1, \dots, x_k}(S, t)$ be given by (6.3) with $d_{x_1, \dots, x_k}(f, g)$ in place of d .

If $N_{x_1, \dots, x_k}(S, t)$ is small, this means that there are not many functions in S which take very different values at points x_1, x_2, \dots, x_k . But maybe the functions in S differ a lot at some other points. To have an adequate measure of how “large” S is, we should select k and points x_1, \dots, x_k such that $N_{x_1, \dots, x_k}(S, t)$ is the maximal possible. The quantity

$$D(S, t) := \ln \left(\sup_k \sup_{x_1, \dots, x_k} N_{x_1, \dots, x_k}(S, t) \right)$$

is called the Koltchinskii–Pollard entropy, or just *metric entropy* of the set S . The larger the metric entropy of S , the more “different”, or “well-separated” functions it contains.

Measuring “Oscillation” and Combinatorial Dimension

In other applications, we would like S to contain functions which “oscillate” well. For example, given $t > 0$, do there exist points $x_1, x_2 \in \Omega$, real numbers h_1, h_2 , and four functions $f_1, f_2, f_3, f_4 \in S$ such that (i) $f_1(x_1) \leq h_1, f_1(x_2) \leq h_2$; (ii) $f_2(x_1) \leq h_1, f_2(x_2) \geq h_2 + t$; (iii) $f_3(x_1) \geq h_1 + t, f_3(x_2) \leq h_2$; and (iv) $f_4(x_1) \geq h_1 + t, f_4(x_2) \geq h_2 + t$, see Fig. 6.6c. In other words, can we find four functions such that one is small at both points, one is large at both, one is small at the first point but large at the second, and one is the other way around? More generally, can we find k points $x_1, \dots, x_k \in \Omega$, and real numbers h_1, \dots, h_k , such that, in whatever way we divide points x_1, \dots, x_k into two groups, there is a function $f \in S$ such that $f(x_i) \leq h_i$ for points in the first group, but $f(x_i) \geq h_i + t$ for points in the other one? The largest k for which this is possible is called the *combinatorial dimension* of the set S and is denoted by $v(S, t)$. It measures how “rich” the set S is, how many “differently oscillating” functions it contains.

The Connection Between Metric Entropy and Combinatorial Dimension

The theorem below, proved in [330], relates combinatorial dimension and metric entropy, and (assuming a minor additional condition on $v(S, t)$) states that these two fundamental quantities are “essentially equivalent up to a constant factor”.

Theorem 6.6 *Let S be a set of functions, and assume that there exists an $a > 1$ such that $v(S, at) \leq \frac{1}{2}v(S, t)$, $\forall t > 0$. Then for all $t > 0$,*

$$c \cdot v(S, 2t) \leq D(S, t) \leq C \cdot v(S, ct),$$

where $c > 0$ is an absolute constant, and C depends only on a .

Before 2006, there were some important open questions about $v(S, t)$, such that the “corresponding” results for $D(S, t)$ were known. With Theorem 6.6, one can immediately derive results about $v(S, t)$ from those about $D(S, t)$ and vice versa. Also, the fact that two fundamentally different ways to measure “how big the set of functions is” turned out to be (essentially) equivalent is a hint that we have found the “correct” method for doing this.

Reference

M. Rudelson and R. Vershynin, Combinatorics of random processes and sections of convex bodies, *Annals of Mathematics* **164**-2, (2006), 603–648.

6.7 On the Approximation of Real Numbers by Irreducible Fractions

Irreducible Fractions

The area of a unit square $ABCD$ is equal to exactly 1. The area of the triangle ABC is exactly $\frac{1}{2}$. Of course, it may also be written down as $\frac{2}{4}$, or $\frac{3}{6}$, or even $\frac{100}{200}$, but $\frac{1}{2}$ is the only way to write down this number in the form of an *irreducible fraction*. A fraction $\frac{a}{b}$, with a, b integers, is called irreducible if the numbers a, b are *coprime*, that is, they have no common factor except for 1 (and -1). For example, the fraction $\frac{6}{12}$ is not irreducible, because $a = 6$ and $b = 12$ have common divisor 3, which can be “crossed out”: $\frac{6}{12} = \frac{2 \cdot 3}{4 \cdot 3} = \frac{2}{4}$. The resulting fraction $\frac{2}{4}$ is again not irreducible because of the common factor 2. After crossing it out, we get an irreducible fraction $\frac{1}{2}$. In a similar way, any fraction $\frac{a}{b}$ can be simplified to an irreducible form.

However, the length $|AC|$ of the diagonal AC of our unit square cannot be written as a fraction $\frac{a}{b}$ for any integers a, b . Indeed, by Pythagoras’ theorem, $|AC|^2 = |AB|^2 + |BC|^2 = 1^2 + 1^2 = 2$, hence $|AC| = \sqrt{2}$. Assume that $\sqrt{2} = \frac{a}{b}$, where $\frac{a}{b}$ is an irreducible fraction. Then $b^2 = 2a^2$, hence b is even and can be written as $b = 2k$. Then $(2k)^2 = 2a^2$, or $2k^2 = a^2$, hence a is even as well. But then a and

b are both divisible by 2, which contradicts the assumption that the fraction $\frac{a}{b}$ is irreducible.

Dirichlet's Approximation Theorem

Numbers of the form $\frac{a}{b}$ for integers a, b are called *rational*, while numbers like $\sqrt{2}$ are called *irrational*. Because there is no way to write down an irrational number x as $\frac{a}{b}$ exactly, a central question is how well it can be approximated as a fraction with given denominator b . For example, if $x = \sqrt{2}$ and $b = 10$, then $bx = 10\sqrt{2}$ belongs to the interval $(14, 15)$, and the closest integer to it (which in this case is $a = 14$) must be at distance less than $\frac{1}{2}$ from bx , see Fig. 6.7a. The inequality $|bx - a| < \frac{1}{2}$ implies that $|x - \frac{a}{b}| < \frac{1}{2b}$.

Can this error estimate be improved? For a given fixed b , it cannot, because there may be irrational numbers which are very close to the mid-point $\frac{a+1/2}{b}$ of the interval $[\frac{a}{b}, \frac{a+1}{b}]$. But, if we are so unlucky, why not try a different denominator? The celebrated Dirichlet approximation theorem states that, for any irrational number x , there are infinitely many fractions $\frac{a}{b}$, such that

$$\left| x - \frac{a}{b} \right| < \frac{1}{b^2}. \quad (6.4)$$

The proof is so simple and elegant that it can be presented here in full. For any real number y , let $[y]$ be the largest integer not exceeding y . Then the difference $y - [y]$, sometimes denoted $\{y\}$, is called the *fractional part* of y . Obviously, $0 \leq \{y\} < 1$ for any y . Now, for a given irrational number x , and positive integer B , consider the sequence of numbers $0, \{x\}, \{2x\}, \dots, \{Bx\}$. Figure 6.7b depicts such a sequence for $x = \sqrt{2}$ and $B = 10$. Because there are $B + 1$ terms in this sequence, and all of them belong to the interval $[0, 1)$, the two closest ones, say $\{nx\}$ and $\{mx\}$, should be at distance less than $\frac{1}{B}$ from each other. But then their difference $|(n - m)x|$ is at distance less than $\frac{1}{B}$ from some integer a , and, with $b = |n - m|$, we get $|bx - a| < \frac{1}{B} \leq \frac{1}{b}$, which implies (6.4).

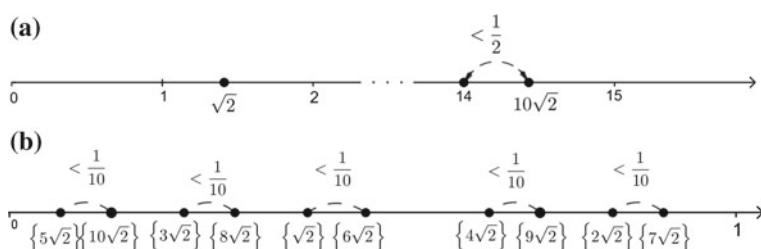


Fig. 6.7 An illustration of the proof of Dirichlet's approximation theorem

A Better Approximation for “Most” Irrational Numbers: Khinchin’s Theorem

The error bound $\frac{1}{b^2}$ in (6.4) can be improved to $\frac{1}{\sqrt{5}b^2}$, but no further: there are infinitely many irrational numbers, for example, the “golden ratio” $\phi = \frac{1+\sqrt{5}}{2}$, for which this bound is the best possible. However, mathematicians were able to prove that such numbers are very “special”, and “almost all” irrational numbers *can* be better approximated!

How can we judge that one set containing infinitely many numbers is “larger” than another one? First, let us consider a simple example: let A be the set of numbers of the form $\frac{1}{2^k}$, $k = 1, 2, 3, \dots$, and let B be the set of all other real numbers in the interval $(0, 1)$. Then both sets A and B are infinite, but, intuitively, the set B is much “larger”. This intuition can be formalized with the notion of *Lebesgue measure*, which, intuitively, is just the “total length” of a set. The set B is the union of intervals $(\frac{1}{4}, \frac{1}{2})$, $(\frac{1}{8}, \frac{1}{4})$, \dots with lengths $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, \dots , respectively, hence the total length of B is the infinite sum $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots + \frac{1}{2^n} + \dots$. One can prove by induction that the sum of the first n terms in this sum is $S_n = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots + \frac{1}{2^n} = 1 - \frac{1}{2^n}$. The value of an infinite sum $\sum_{n=1}^{\infty} a_n$ is defined as the limiting value of S_n as n goes to infinity, so in this case it is equal to 1. But the length of the full interval $(0, 1)$ is 1 as well, hence the length of A should be $1 - 1 = 0$. Formally, any set of real numbers has Lebesgue measure 0 if it can be fully covered by intervals of total length ε for any $\varepsilon > 0$. If a property holds for all real numbers, except, possibly, for a set of measure 0, we say it holds for *almost every* real number.

It follows from Khinchin’s theorem [222], proved in 1926, that, for every $\varepsilon > 0$, and almost every real number x , there are infinitely many fractions $\frac{a}{b}$ such that

$$\left| x - \frac{a}{b} \right| < \frac{g(b)}{b}, \quad (6.5)$$

where g is a function such that $\sum_{b=1}^{\infty} g(b) = \infty$ —in other words, for any C , no matter how large, there is a B such that $\sum_{b=1}^B g(b) > C$. For example, it is known that this property holds for $g(b) = \frac{\varepsilon}{b}$ for any $\varepsilon > 0$, hence $|x - \frac{a}{b}| < \frac{\varepsilon}{b^2}$ holds true for almost every x , and infinitely many fractions $\frac{a}{b}$.

Approximation by Irreducible Fractions: The Duffin–Schaeffer Conjecture

However, Khinchin’s theorem allows us to count the same approximation many times with different values of b . For example, $\frac{14}{10}$ approximates $\sqrt{2}$ with $b = 10$, while $\frac{7}{5}$ approximates it with $b = 5$. However, in fact $\frac{14}{10}$ and $\frac{7}{5}$ is just the same fraction written differently, so it should not be counted twice. So, let us require that a and b in (6.5) are coprime, so that $\frac{a}{b}$ is an irreducible fraction.

Duffin and Schaeffer [120] conjectured in 1941 that, for almost every x , (6.5) has infinitely many *coprime* solutions a, b , if and only if $\sum_{b=1}^{\infty} g(b) \frac{\phi(b)}{b} = \infty$, where $\phi(b)$ is the number of integers which are less than b and co-prime with it. For example, there are two integers (1 and 5) which are less than 6 and coprime with 6, hence $\phi(6) = 2$. The Duffin–Schaeffer conjecture is a natural generalization of Khinchin’s theorem, but no-one knows how to prove it.

A More General Conjecture

However, even the Duffin–Schaeffer conjecture is not general and powerful enough to completely answer the question about coprime solutions to (6.5). The main problem is that it is formulated for “almost every” x , where “almost every” is defined using the Lebesgue measure. In fact, Lebesgue measure is just one special way to measure the “size” of a set of real numbers. A much more general way is the following. Take any function $f : [0, \infty) \rightarrow [0, \infty)$, which is continuous, non-decreasing, and $f(0) = 0$, for example, $f(x) = x$ or $f(x) = x^2$ works. Then, given a set S of real numbers, and any $\delta > 0$, try to cover S by intervals of length $r_1, r_2, \dots, r_n, \dots$ such that $r_i \leq \delta$ for all i and the sum $\sum_{i=1}^{\infty} f(r_i/2)$ is as small as possible. Let $H_{\delta}^f(S)$ be the smallest value of this sum, and $H^f(S) = \lim_{\delta \rightarrow 0} H_{\delta}^f(S)$. The number $H^f(S)$ is called the Hausdorff f -measure of the set S . Now, let us say that some property holds for f -almost every real number in an interval $[c, d]$ if it holds for some $S \subset [c, d]$ such that $H^f(S) = H^f([c, d])$. This is a very general and powerful way to formalize the intuition that a property holds for almost all real numbers, except for some very special counterexamples.

In 2006, Beresnevich and Velani [41] argued that, if we are really interested in solving (6.5) for “almost every” x , we should use the general definition of “almost every”, and conjectured that for f -almost every x , (6.5) has infinitely many coprime solutions a, b , if and only if $\sum_{b=1}^{\infty} f(\frac{g(b)}{b}) \phi(b) = \infty$.

The Special Case Implies the General One!

Of course, the Duffin–Schaeffer conjecture is a very special case of the Beresnevich–Velani conjecture. However, the following theorem states that the full general conjecture can be deduced from this very special case!

Theorem 6.7 *The Duffin–Schaeffer conjecture implies the Beresnevich–Velani conjecture.*

With Theorem 6.7, it now “remains” to prove the original Duffin–Schaeffer conjecture to answer the question in full generality. The “only” problem is that no-one has been able to do this since 1941...

Reference

V. Beresnevich and S. Velani, A Mass Transference Principle and the Duffin–Schaeffer conjecture for Hausdorff measures, *Annals of Mathematics* **164**–3, (2006), 971–992.

Chapter 7

Theorems of 2007



7.1 Bounding the Error in Approximation of Smooth Functions by Polynomials

From Temperature Measurement to Approximation Theory

Assume that you need to measure and report how the outside temperature changes from 6 AM to 10 AM. How would you measure it and how would you report the result? First, it is clear that the temperature cannot be measured exactly, and the best you can do is to report its approximate value. Second, to measure it at every moment of time, you would need to perform infinitely many measurements, and report infinitely many data.

Instead, you could perform measurements at some pre-specified times, say, at 6.00, 6.05, and so on, up to 10.00, plot the data on the graph with x -axis and y -axis representing time and temperature, respectively, and, based on these data, try to guess the approximate value of the temperature at any time x when you did not perform the measurement. For example, if your data graphically resembles a straight line with equation $y = ax + b$, then you can guess that in fact at *every* time x the temperature was approximately $ax + b$.

This is the topic of study of the mathematical field called approximation theory. We have some complicated function $f(x)$, so difficult that it has no exact formula for general x , and the best we can do is to calculate the value of f at some specific points x_k . The problem is to approximate it by some “simpler” function g , that is, find a function g defined by some simple formula such that $f(x) \approx g(x)$ for all x belonging to some interval. In our example above, the function $f(x)$ representing the temperature at time x was approximated by a linear function $g(x) = ax + b$.

Uniform “Points for Measurements” and Polynomial Approximation

In general, assume that the function $f(x)$ is defined on some interval, say $[-1, 1]$, and that we have decided to perform N measurements/calculations and calculate values $f(x_1), f(x_2) \dots f(x_N)$ at some points $-1 < x_1 < x_2 < \dots < x_N < 1$. Where should we select the points? If, for example, x_1, \dots, x_N are all positive, then the measurements would give us no information about $f(x)$ for $x \in [-1, 0]$. It is intuitively obvious that it is better to select points which cover $[-1, 1]$ as uniformly as possible, so let us select

$$x_1 = -\frac{N-1}{N}, x_2 = -\frac{N-3}{N}, \dots, x_k = -1 + \frac{2k-1}{N}, \dots, x_{N-1} = \frac{N-3}{N}, x_N = \frac{N-1}{N}.$$

For example, for $N = 5$ the selected points would be $-\frac{4}{5}, -\frac{2}{5}, 0, \frac{2}{5}, \frac{4}{5}$.

The next step is to select a simple function g such that $f(x_k) \approx g(x_k)$ for all $k = 1, \dots, N$. More formally, let $\varepsilon(f, g)$ be the smallest real number such that

$$|f(x_k) - g(x_k)| \leq \varepsilon(f, g), \quad k = 1, 2, \dots, N.$$

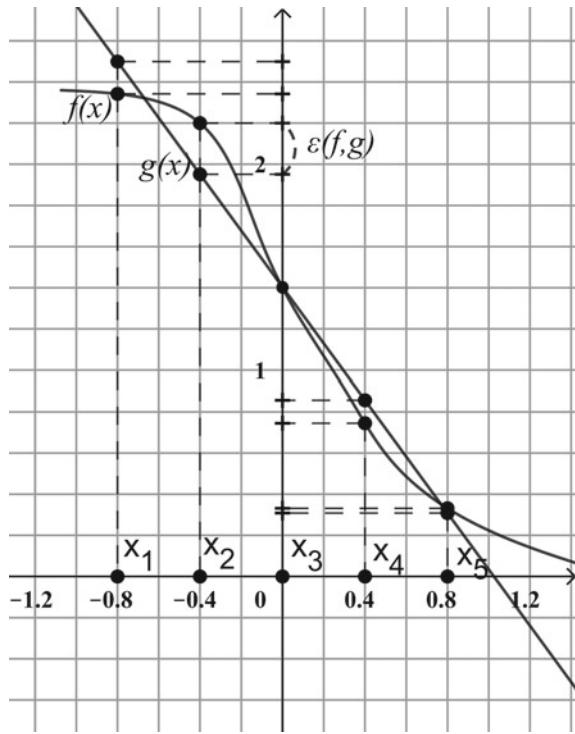
In other words, $\varepsilon(f, g) = \max_k |f(x_k) - g(x_k)|$. Of course, the smaller $\varepsilon(f, g)$, the better the approximation. For example, we can try a linear function $g(x) = ax + b$ and find coefficients a, b such that $\varepsilon(f, ax + b)$ is as small as possible, see Fig. 7.1. If a^*, b^* are optimal coefficients, and $\varepsilon(f, a^*x + b^*)$ is, for example, 0.01, we may be satisfied with the quality of approximation. However, if $\varepsilon(f, a^*x + b^*)$ is large, we may decide that no linear function approximates the data well enough, and try, for example, a quadratic polynomial $g(x) = ax^2 + bx + c$. Then, again, we find a, b, c such that $\varepsilon(f, ax^2 + bx + c)$ is as small as possible, and, if we are still not happy, try g in the form of a cubic polynomial, and so on.

Can it be that we could try g in the form of a linear, quadratic, cubic polynomial, and so on up to infinity, but never get a small $\varepsilon(f, g)$? No, it cannot be. In fact, if g is a polynomial of degree $n = N - 1$, we can always guarantee that $\varepsilon(f, g) = 0$, that is, $g(x_k)$ is exactly equal to $f(x_k)$ for $k = 1, 2, \dots, N$. The reason is that the equalities $f(x_k) = g(x_k)$, $k = 1, 2, \dots, N$ form a system of N equations, the coefficients of g are N unknowns, and one can prove that this system always has a solution. Of course, going to a polynomial of degree as large as $N - 1$ is the worst-case scenario: in practice, we would hope that some polynomial g with much smaller degree n provides a reasonably good approximation.

Controlling the Quality of Approximation on the Whole Interval

The final step is that we claim: ok, because $f(x_k) \approx g(x_k)$ for all $k = 1, \dots, N$, then we can assume that actually $f(x) \approx g(x)$ for all $x \in [-1, 1]$, and report $g(x)$ as a “good” approximation to f on the *whole* interval.

Fig. 7.1 Approximation by a linear function



Of course, this last step is not always well-justified. In general, f may be close to some “nice” function g for $x = x_k$, $k = 1, \dots, N$, but behave in a different way “in between”. For example, imagine a function f such that $f(x) = 0$ for all rational x but $f(x) = 1$ for irrational x . Then, with $g(x) = 0$, the approximation $f(x) \approx g(x)$ works perfectly well for $x = x_k$, $k = 1, \dots, N$, because all x_k are rational, but it is far from being correct for all $x \in [-1, 1]$.

However, we may hope to rigorously justify the correctness of the final step above at least if f is “sufficiently smooth”. To formulate the problem rigorously, let us define the ratio

$$K_{f,g,N}(x) = \frac{|f(x) - g(x)|}{\varepsilon(f, g)} = \frac{|f(x) - g(x)|}{\max_k |f(x_k) - g(x_k)|}.$$

For example, if we could prove that $K_{f,g,N}(x) \leq 2$, this would mean that $|f(x_k) - g(x_k)| \leq \varepsilon$, $k = 1, \dots, N$ implies $|f(x) - g(x)| \leq 2\varepsilon$ for all $x \in [-1, 1]$. The problem is then to derive some reasonably good upper bounds for $K_{f,g,N}(x)$.

Some Lower and Upper Bounds

Because the general problem is very difficult, researchers have tried to prove some rigorous bounds on $K_{f,g,N}(x)$ at least in the “trivial” case $f(x) = 0$. Let $K_{N,n}(x)$ be the smallest real number such that the bound

$$K_{0,g,N}(x) \leq K_{N,n}(x)$$

holds for all polynomials g of degree n . In other words, $K_{N,n}(x) = \max_{g \in G_n} K_{0,g,N}(x)$, where G_n is the set of all polynomials of degree n . In 1992, Coppersmith and Rivlin [103] proved that

$$e^{c_1 n^2/N} \leq \max_{x \in [-1, 1]} K_{N,n}(x) \leq e^{c_2 n^2/N}. \quad (7.1)$$

The second inequality in (7.1) provides a good upper bound for $K_{N,n}(x)$ for all $x \in [-1, 1]$, provided that n is less than \sqrt{N} . However, as we discussed above, there is no warranty that we can find a good approximation to any function by a polynomial of degree $n < \sqrt{N}$, in fact, the worst-case scenario is $n = N - 1$. But in this case the first inequality implies that $\max_{x \in [-1, 1]} K_{N,n}(x)$ goes to infinity if $N \rightarrow \infty$, so there is no hope to bound $K_{N,n}(x)$ for all $n < N$ and all $x \in [-1, 1]$.

A Better Bound on a Subinterval

The following theorem of Rakhmanov [312] proves, however, that such a bound is possible at least inside the subinterval $x \in (-r, r)$, where

$$r = \sqrt{1 - n^2/N^2}. \quad (7.2)$$

Theorem 7.1 *With r defined as in (7.2) and $n < N$,*

$$K_{N,n}(x) \leq C \ln \frac{\pi}{\arctan \left(\frac{N}{n} \sqrt{r^2 - x^2} \right)}, \quad \forall x \in (-r, r),$$

where C is an absolute constant.

In case you do not know, $\arctan(y)$ for $y > 0$ denotes the size, measured in radians, of angle A in a right triangle ABC with right angle C and lengths of legs $|AC| = 1$ and $|BC| = y$. Theorem 7.1 can be used to show that, when you approximate a “smooth” f by a polynomial of degree $n < N$, and your approximation works well for $x = x_k$, $k = 1, \dots, N$, then it is guaranteed to work well for all x belonging to a closed subinterval of $(-r, r)$. Rakhmanov also showed that $(-r, r)$ is the maximal subinterval of $[-1, 1]$ with this property.

Reference

E. Rakhmanov, Bounds for polynomials with a unit discrete norm, *Annals of Mathematics* **165**-1, (2007), 55–88.

7.2 Superlinear Growth of Digit Patterns in the Decimal Expansion of Algebraic Numbers

Decimal Expansions of Rational and Irrational Numbers

Numbers which can be written as the ratio $\frac{a}{b}$ of two integers a, b are called *rational numbers*. Some rational numbers, like $\frac{1}{2} = 0.5$, have a finite decimal expansion, while the expansion of others, like $\frac{3}{22} = 0.136363636363\ldots$ is infinite, but still very simple. In this example, it starts with 0.1, followed by a “period” 36, which repeats itself infinitely often. Such an expansion is called “eventually periodic”. It is easy to prove that in fact the decimal expansion of every rational number is eventually periodic, and, vice versa, every real number with eventually periodic decimal expansion is rational.

It has been known since ancient times that not all real numbers are rational. For example, $\sqrt{2}$ cannot be written as a ratio $\frac{a}{b}$ of integers, and its decimal expansion $\sqrt{2} = 1.414213562373\ldots$ does not repeat itself in a periodic way. Instead, the digits in it seem to appear in an unpredictable, chaotic way.

Some “Patterns” in Decimal Expansions

The decimal expansion of $\frac{3}{22}$ contains digits 0, 1, 3, and 6, and only 3 and 6 appear infinitely often. In contrast, the decimal expansion of $\sqrt{2}$ contains all digits, and we expect all of them to appear infinitely often. There are no reasons why some digit, say 7, would appear just finitely many times, and then stop appearing at some moment.

Similarly, in $\frac{3}{22} = 0.136363636363\ldots$ we see just four two-digit combinations: 01, 13, 36, and 63, of which 01 and 13 appear just once. In contrast, the decimal expansion of $\sqrt{2} = 1.414213562373\ldots$ contains combinations 14, 41, 14 again, 42, 21, 13, and so on, in fact it contains all $10^2 = 100$ possible combinations, and it is conjectured that all of them should appear infinitely often.

Let $p(n)$ be the number of distinct combinations/blocks of length n occurring in the decimal expansion of a given number: that is, $p(1)$ is the number of distinct digits, $p(2)$ in the number of occurring blocks of 2 digits, and so on. Then, for $\frac{3}{22}$, $p(1) = 4$, $p(2) = 4$ (blocks 01, 13, 36, 63), $p(3) = 4$ (blocks 013, 136, 363, 636), and so on, $p(n) = 4$ for all n . The same is true for any rational number: there is a bound B such that $p(n) \leq B$ for all n .

For $\sqrt{2}$, we expect *all* possible blocks to occur, that is, $p(1) = 10$ (all 10 digits occur), $p(2) = 10^2 = 100$ (all 100 combinations from 00 to 99), $p(3) = 10^3 = 1000$, and we expect that $p(n) = 10^n$ for all n .

A General Lower Bound for $p(n)$

The conjecture “ $p(n) = 10^n$ for $\sqrt{2}$ ” is very far from being proved, but can we at least establish some lower bounds for $p(n)$? First, it is relatively easy to prove that $p(n) \geq n + 1$ for every irrational number n . However, in general this bound cannot be improved. Consider sequences of 0’s and 1’s made with the following rule: $S_0 = 0$, $S_1 = 01$, $S_2 = S_1 \cup S_0 = 010$, $S_3 = S_2 \cup S_1 = 01001$, $S_4 = S_3 \cup S_2 = 01001010$, $S_5 = S_4 \cup S_3 = 0100101001001$, and so on, $S_n = S_{n-1} \cup S_{n-2}$ for all $n \geq 2$. Let S be the infinite sequence made by this rule, and let

$$\alpha = 0.0100101001001\dots$$

be an irrational number whose decimal expansion is zero, then point, and then the sequence S . Then, for α , $p(1) = 2$ (only digits 0 and 1 are used), $p(2) = 3$ (only three blocks 00, 01, and 10 occur), $p(3) = 4$ (blocks 001, 010, 100, and 101), and so on, one can prove that $p(n) = n + 1$ for every n . Hence, the bound $p(n) \geq n + 1$ is the best possible lower bound valid for *all* irrational numbers. To make further progress for $\sqrt{2}$, we need to use some special properties it has.

Better Bounds for Algebraic Numbers

Of course, the special property of $x = \sqrt{2}$ is that $x^2 = 2$. In other words, $\sqrt{2}$ is a solution to the equation $x^2 - 2 = 0$. Real numbers which are solutions to equations with rational coefficients are called *algebraic numbers*. For example, $x = \sqrt[5]{1.5}$ is also an algebraic number, because it is a solution to the equation $2x^5 - 3 = 0$. In 1997, Ferenczi and Mauduit [150] proved that no algebraic number can have $p(n) = n + 1$ for all n . In particular, this implies that the number α defined above is *not* a solution to any equation with rational coefficients. Such numbers are called *transcendental*.

The result of Ferenczi and Mauduit implies that for every irrational algebraic number we have $p(n) > n + 1$ for sufficiently large n . That is, we can have $p(1) = 2$, $p(2) = 3$, but then, for some natural number N_1 , we should have $p(N_1) > N_1 + 1$, and then the inequality $p(n) > n + 1$ continues to hold for all $n \geq N_1$. Moreover, the same is true for every constant c in place of 1: whatever large c we choose, there exists a natural number N_c such that $p(n) > n + c$ for all $n \geq N_c$.

The bound $p(n) > n + c$ is better than nothing, but it is very far away from the conjecture that $p(n) = 10^n$. For example, it could be that $p(n)$ for $\sqrt{2}$ satisfies $p(n) \approx 1.01n$ for large n . This would not contradict the Ferenczi–Mauduit result, because, for $n > 100c$, $1.01n > n + c$.

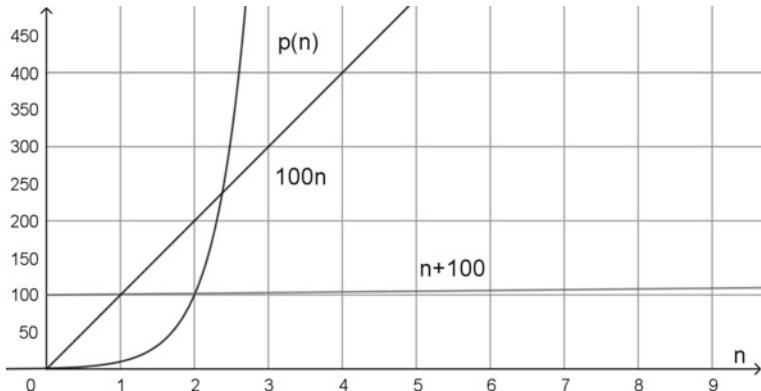


Fig. 7.2 Conjectural growth of $p(n)$, together with old and new lower bounds, with $c = 100$

A Dramatic Improvement

The following theorem of Adamczewski and Bugeaud [4] provides a much better bound.

Theorem 7.2 *For every irrational algebraic number β , and every constant c , there is a natural number N such that*

$$p(n) > c \cdot n, \quad \forall n \geq N.$$

In other words,

$$\lim_{n \rightarrow \infty} \frac{p(n)}{n} = +\infty.$$

Theorem 7.2 implies that whatever large c we choose – say, a million – then, for n large enough, we will be able to find more than a million times n distinct digit blocks of length n in the decimal expansion of $\sqrt{2}$. And the same is true if $\sqrt{2}$ is replaced by $\sqrt[5]{1.5}$, or any solution of any equation with rational coefficients. However, Theorem 7.2 is still not strong enough to guarantee that the decimal expansion of $\sqrt{2}$ contains, for example, the digit 7 infinitely often. Even without using 7 at all we could form 9^n different blocks of length n , so even the bound $p(n) \geq 9^n$ would not suffice for this, while Theorem 7.2 only gives $p(n) > c \cdot n$. However, it is still a huge improvement on the previous bound $p(n) > n + c$, see Fig. 7.2, and we now have more reasons to be optimistic about further progress in this very interesting research direction.

Extension to Different Number Systems

So far, we have defined $p(n)$ as the number of distinct blocks of length n in the decimal expansion of x . However, there is nothing special about the number 10, it is

just a historical coincidence that we use the decimal number system. For every integer $b \geq 2$, we can write all numbers in the number system with base b instead of 10. For example, with $b = 5$, numbers one, two, three, four, five, six, seven, eight, nine, ten, eleven, and so on would be written as 1, 2, 3, 4, 10, 11, 12, 13, 14, 20, 21, In the usual number system with base $b = 10$, eleven is written as 11 just because $11 = 1 \cdot 10 + 1 = 1 \cdot b + 1$, and one hundred and twenty is written as 120 because it is equal to $1 \cdot b^2 + 2 \cdot b + 0$. In the number system with $b = 5$, eleven is written as 21 because it is equal to $2 \cdot 5 + 1 = 2 \cdot b + 1$, while one hundred and twenty, which is equal to $4b^2 + 4 \cdot b + 0$, should be written as 440. The system with $b = 2$ is called binary, the first natural numbers in it look like 1, 10, 11, 100, 101, 110, 111, 1000, 1001, and so on. A hundred and twenty here is written as 1111000 because it is equal to $1 \cdot b^6 + 1 \cdot b^5 + 1 \cdot b^4 + 1 \cdot b^3 + 01 \cdot b^2 + 0 \cdot b + 0$. The binary system is used for representing numbers for computers. All real numbers like $\sqrt{2}$ can be represented as an infinite sequence of digits in any number system with any base $b \geq 2$. Then $p(n)$ can be defined in exactly the same way as the number of distinct blocks of length n is such an expansion, and Theorem 7.2 remains correct in this generality.

Reference

B. Adamczewski and Y. Bugeaud, On the complexity of algebraic numbers I. Expansions in integer bases, *Annals of Mathematics* **165**-2, (2007), 547–565.

7.3 The Existence of Intervals with Too Many and Too Few Primes

Hidden Mysteries in the Sequence of Primes

Primes, that is, natural numbers having exactly two divisors, 1 and themselves, have been studied in mathematics since ancient times, but are still surrounded by many secrets, mysteries, and surprises. For example, we still do not know if there are infinitely many primes p such that $p + 2$ is also a prime (such pairs are called *twin primes*), and this is one of the oldest open questions in the whole of mathematics.

One of the oldest theorems in mathematics is Euclid's proof that there are infinitely many primes. By contradiction, assume that there are finitely many, and they are p_1, p_2, \dots, p_n . Then the number $N = p_1 \cdot p_2 \cdots \cdot p_n + 1$ is not divisible by any of p_1, \dots, p_n , and therefore should either be a prime, or should be divisible by some other prime p_{n+1} – in any case, a contradiction.

Ok, there are infinitely many primes, but how many of them would we expect to be within the first, say, $x = 10^{1000}$ integers? Let $\pi(x)$ be the number of primes less than x . Because the first primes are 2, 3, 5, 7, 11, 13, 17, 19, 23, ..., $\pi(10) = 4$, and $\pi(20) = 8$. The famous prime number theorem states that, for large x , $\pi(x)$ is approximately equal to $x / \ln x$, where \ln is the natural logarithm, see Fig. 7.3. For $x = 10^{1000}$, $\ln(x) \approx 2302.6$, that is, among the integers with at most 1000 digits, approximately every 2300th is prime. In other words, if we select such an integer at

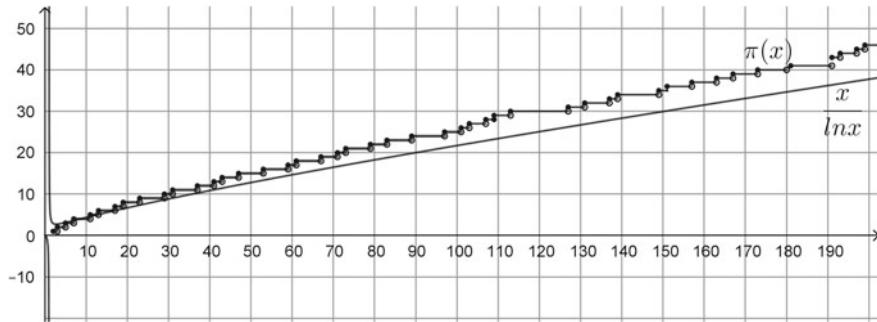


Fig. 7.3 Graphs of $\pi(x)$ and $x / \ln x$

random, the probability that it will be a prime is about $1/2300$. In general, a random integer less than x is prime with probability about $1/\ln x$.

An Informal Argument for the Twin Primes Conjecture

Because many natural questions about primes are so difficult, people try to “guess” the answer using the following method. Take a 1000-digit integer at random, with probability about $1/2300$ it is prime, denote it p . Then check if $p + 2$ is prime as well, with probability $1/2300$ it should be. Thus, with probability about $(1/2300)^2$ both p and $p + 2$ are primes. This means that there should be about $10^{1000}/2300^2$ pairs of twin primes less than 10^{1000} , or, more generally, about $x/(\ln x)^2$ pairs of twin primes less than x . Because the function $f(x) = x/(\ln x)^2$ goes to infinity as x increases, there should be infinitely many twin primes.

Of course, one should be careful with this type of argument. For example, one could argue in exactly the same way that there should be infinitely many primes p such that $p + 1$ is prime as well, while in reality $p = 2$ is the only prime with this property. This is because either p or $p + 1$ is even, and the only even prime number is 2. In the argument above, if p happened to be prime, then almost surely p is odd, hence $p + 1$ is even and cannot be prime. In contrast, $p + 2$ is odd, which increases its chances to be prime. Similarly, if p is prime, it is not divisible by 3 (unless $p = 3$), hence $p + 2$ cannot give remainder 2 when divided by 3, it can either give remainder 0 or 1. Thus, the probability that $p + 2$ is divisible by 3 is about $1/2$ instead of $1/3$. Continuing this way, we find that the probability that $p + 2$ is prime is not really $1/2300$, but is $C/2300$ for some constant C which can be explicitly computed. This, however, does not change the conclusion: because the function $f(x) = Cx/(\ln x)^2$ goes to infinity as x increases, we expect that there should be infinitely many twin primes. Subject to this “divisibility correction”, the above informal argument, while not a formal proof, is often considered to be very convincing.

Counting Primes in Intervals

However, we next tell a story of when such an argument failed in a surprising way. For this, consider a very natural question: how many primes would we expect to be in the interval $[x + 1, x + y]$? Well, there are y integers there, each integer $p \in [x + 1, x + y]$ is prime with probability about $1/\ln p$. If y is much less than x , then $1/\ln p \approx 1/\ln x$. Hence, in total we would expect about $y/\ln x$ primes in this interval. If y is very small, say $y = 1$ or $y = 2$, the “divisibility” problem arises in the same way as above. However, we may choose y large but still much less than x , for example, $y = (\ln x)^A$ for some constant $A \geq 1$, and then the heuristic above, intuitively, should work. Using the notation π defined above, we expect

$$\pi(x + y) - \pi(x) \approx \frac{y}{\ln x}, \quad y = (\ln x)^A, \quad A \geq 1.$$

Or, making the meaning of \approx more precise, one could conjecture that

$$\lim_{x \rightarrow \infty} \frac{\pi(x + y) - \pi(x)}{y/\ln x} = \lim_{x \rightarrow \infty} \frac{\pi(x + (\ln x)^A) - \pi(x)}{(\ln x)^{A-1}} = 1.$$

The Failure of the Informal Argument

Surprisingly, Maier [259] showed in 1985 that this conjecture turned out to be false! There are intervals $[x + 1, x + y] = [x + 1, x + (\ln x)^A]$ containing significantly more primes than they “should”, while there are other intervals containing significantly fewer primes! More formally, there is a constant $\delta_A > 0$, depending on A , such that $\pi(x + y) - \pi(x) > (1 + \delta_A) \frac{y}{\ln x}$ for infinitely many values of x , and $\pi(x + y) - \pi(x) < (1 - \delta_A) \frac{y}{\ln x}$ for some other, but also infinitely many, values of x .

Of course, one can ask many further questions about Maier’s result. For example, for exactly which pairs (x, y) do the “Maier type irregularities” happen, and how large are these “irregularities”? In 2007, Granville and Soundararajan [173] proved a very general theorem, which they described as “perhaps the best possible result in this context”.

Intervals with Too Many and Too Few Primes

To formulate the theorem we need some preparation. First, it is convenient to count each prime p with “weight” $\ln p$, that is, study the function

$$\theta(x) = \ln 2 + \ln 3 + \ln 5 + \cdots + \ln p_x = \sum_{p < x} \ln p,$$

where p_x is the largest prime less than x . Alternatively, $\theta(x)$ is the logarithm of the product of all primes less than x . Then the prime number theorem states that $\theta(x) \approx x$ for large x , hence the number of “weighted” primes on $[x+1, x+y]$ is about $\theta(x+y) - \theta(x) \approx y$. The “quality” of this approximation can be measured by

$$\Delta(x, y) := \frac{\theta(x+y) - \theta(x) - y}{y}.$$

If $\Delta(x, y)$ is large and positive, then $[x+1, x+y]$ contains much more primes than it “should”, and if large and negative – much fewer. The following theorem of Granville and Soundararajan provides conditions under which both “irregularities” are guaranteed to happen.

Theorem 7.3 *Let x be large and y be such that*

$$\ln x \leq y \leq \exp\{\beta\sqrt{\ln x}/2\sqrt{\ln \ln x}\},$$

where $\beta > 0$ is an absolute constant. Then there exist numbers x_+ and x_- in $(x, 2x)$ such that

$$\Delta(x_+, y) \geq y^{-\delta(x, y)} \quad \text{and} \quad \Delta(x_-, y) \leq -y^{-\delta(x, y)},$$

where

$$\delta(x, y) = \frac{1}{\ln \ln x} \left(\ln \left(\frac{\ln y}{\ln \ln x} \right) + \ln \ln \left(\frac{\ln y}{\ln \ln x} \right) + O(1) \right).$$

Here, $O(1)$ is some function $f(x, y)$ such that $|f(x, y)| \leq C$ for some constant C . The formulas are a bit complicated, but the idea is simple: the first formula defines for what range of y “irregularities” happen, while the second formula (with $\delta(x, y)$ defined in the third one) establishes “by how much”.

No Arithmetic Sequence is Well-Distributed

In fact, Granville and Soundararajan proved a much more general result stating that “irregularities” (similar to those described in Theorem 7.3) occur not only in the sequence of primes but in fact in *any* sequence of integers that is “arithmetic” in nature. The exact definition of “arithmetic sequence” is too technical to be presented here, but, intuitively, this is any sequence with some arithmetic structure, for example, any subsequence of the sequence of primes is “arithmetic”, so is the sequence of integers representable in the form $x^2 + y^2$, etc. The essence of Granville and Soundararajan’s main theorem is that *no arithmetic sequence is very well-distributed*. This result is really a big surprise, because, if we select a subset of integers at random, then no such “irregularities” occur with probability 1. So, we know now that any set of integers with some arithmetic structure behaves very differently from a random set.

Reference

A. Granville and K. Soundararajan, An uncertainty principle for arithmetic sequences, *Annals of Mathematics* **165**-2, (2007), 593–635.

7.4 Interval Exchange Transformations Are Almost Always Weakly Mixing

How Not to Shuffle a Deck of Cards

Can we build a robot which could do some simple actions, for example, shuffle a deck of cards? The robot would need exact instructions what to do, and the simplest approach would be to teach him some “action”, and then ask to repeat it again and again, until the cards become well-shuffled. The simplest “action” you can try is to take k cards from the bottom and put them on the top of the deck. For example, if there are $n = 5$ cards, which we denote by $(1, 2, 3, 4, 5)$, and $k = 2$, then this “action” would result in moving cards 4 and 5 from the bottom to the top, resulting in the new order $(4, 5, 1, 2, 3)$. Repeating the same “action” results in $(2, 3, 4, 5, 1)$, then $(5, 1, 2, 3, 4)$, then $(3, 4, 5, 1, 2)$, and then we get $(1, 2, 3, 4, 5)$ again. We can see that, with this method, the cards will never be well shuffled. For example, card 1 is always followed by 2, except when 1 is the last card at the bottom, but then 2 is the first one on the top. Card 2 is always followed by 3 with the same exception, and so on. Hence, the described “action”, which is called *rotation*, is too simple to get the deck of cards well-shuffled.

Maybe we could take k cards from the bottom and put them *in the middle* of the deck, say, transform $(1, 2, 3, 4, 5)$ into $(1, 2, 4, 5, 3)$? But this is even worse, because card 1 stays at the top, and will stay there if we repeat this action again and again, no matter how many times. Also, if we take cards from the middle (say 2 and 3) and put them on the top, transforming $(1, 2, 3, 4, 5)$ into $(2, 3, 1, 4, 5)$, then cards 1, 2, 3 stay at the top (although in a different order), and, if we repeat this “action”, we will always get a result in which 1, 2, 3 (in some order) go first, while 4 and 5 stay at the bottom. An “action” such that, for some k , cards $1, 2, \dots, k$ stay at the top, while cards $k + 1, \dots, n$ stay at the bottom, is called *reducible*, while an action which avoids this problem is called *irreducible*.

How to Shuffle a Deck of Cards

Let us try a more complicated “action”, for example, take k cards from the bottom to the top, and at the same time m cards from the top to the bottom. For example, with $k = 2$, $m = 1$, and $n = 5$ cards, the initial order $(1, 2, 3, 4, 5)$ will change to $(4, 5, 2, 3, 1)$. If we repeat it again – move two cards 3, 1 to the top, and one card 4 to

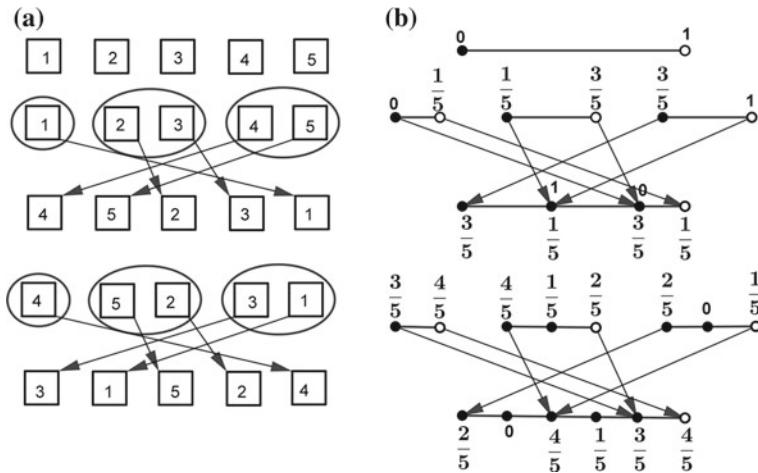


Fig. 7.4 Shuffling **a** a deck of cards, **b** the interval $[0,1]$

the bottom, we get the order $(3, 1, 5, 2, 4)$ – it now looks like the cards are shuffled reasonably well, see Fig. 7.4a.

What we really did here is divided the deck of cards into three groups: m cards on the top (call this group G_1), $n - m - k$ cards in the middle (call it G_2), and k cards on the bottom (group G_3), and then rearranged these groups in the opposite order: G_3 goes first, then G_2 , then G_1 . More generally, we could divide all cards into 4 groups, G_1, G_2, G_3 , and G_4 , and rearrange them somehow, for example, put G_2 first, then G_4 , then G_1 , then G_3 . Even more generally, we can divide cards into d groups and then rearrange. This procedure can be easily “programmed” for the robot. Then we can repeat it again and again, and stop when the cards are “well-shuffled”.

When Can the Cards be Considered to be “Well-Shuffled”?

However, when exactly should the robot stop? Intuitively, it is “clear” that, say, $(3, 4, 5, 1, 2)$ is not a good “shuffle” of $(1, 2, 3, 4, 5)$, while the result $(3, 1, 5, 2, 4)$ looks much “better”. But, for the robot, we would need an exact mathematical criterion for which “shuffles” should be called “good” and which are “not good”.

Let us take standard deck with 52 cards, number them from 1 to 52 in order as they are, then shuffle, look at the first 10 cards, and imagine they are numbered as 15, 11, 26, 20, 13, 1, 14, 21, 4, 12. Would you call such a deck “well-shuffled”? Not really, and the problem is that all these 10 cards are from the region between 1 and 26, while there are no cards from the region between 27 and 52. In a “well-shuffled” deck, the first 10 cards are expected to “cover” the full range from 1 to 52 “more or less uniformly”. In particular, we expect about half of them to be between 1 and 26, and about half between 27 and 52.

Of course, this is true not only for the first 10 cards: if we take $a > 0$ last cards from a “well-shuffled” desk, or a cards from the middle, in positions from $k + 1$ to $k + a$ (denote their numbers as $x_{k+1}, x_{k+2}, \dots, x_{k+a}$), we expect about half of x_i to not exceed 26, and about half to be greater than or equal to 27. By a similar logic, we expect about $\frac{b}{52}$ of the numbers $x_{k+1}, x_{k+2}, \dots, x_{k+a}$ to be between $m + 1$ and $m + b$ whenever $1 \leq m + 1 \leq m + b \leq 52$.

Shuffling Infinitely Many Cards

Can we use the same ideas to teach our robot to shuffle infinitely many “cards”? Specifically, let the “cards” be all real numbers in the interval $[0, 1)$. We would like to develop a procedure to “rearrange” these real number so that they look “well-shuffled”. We can use the same idea as above: first, divide all “cards” into d groups, but this time the groups are intervals

$$G_1 = [0, \lambda_1), G_2 = [\lambda_1, \lambda_1 + \lambda_2), \dots, G_d = \left[\sum_{i=1}^{d-1} \lambda_i, 1 \right),$$

then rearrange/permute these intervals, and then repeat this procedure as many times as needed. For example, the interval $[0, 1)$ may be divided into $d = 3$ intervals $G_1 = [0, 1/5)$, $G_2 = [1/5, 3/5)$, and $G_3 = [3/5, 1)$. Then we rearrange the interval, in, for example, opposite order: G_3, G_2, G_1 . Thus, interval $G_3 = [3/5, 1)$ of length $2/5$ goes to $[0, 2/5)$, then we put G_2 (also of length $2/5$) at the position $[2/5, 4/5)$, and finally interval G_1 (of length $1/5$) goes to $[4/5, 1)$. Such a procedure is called *interval exchange*.

Now, let us repeat it k times, and ask ourselves if the interval $[0, 1)$ is now “well-shuffled”? To decide this, take any interval $A = [x, x + a) \subset [0, 1)$, and consider the set $f^{-k}(A)$ of real numbers which ended up in A after k iterations. If our shuffling is “good”, we would expect the set $f^{-k}(A)$ to be “spread out more or less uniformly” through $[0, 1)$. In particular, we do not expect that $f^{-k}(A) \subset [0, 1/2)$ or $f^{-k}(A) \subset [1/2, 1)$ – instead, we expect that $|f^{-k}(A) \cap (0, 1/2)| \approx |f^{-k}(A) \cap (1/2, 1)| \approx \frac{1}{2}|f^{-k}(A)| = \frac{1}{2}|A|$, where $|X|$ denotes the total length of the set X . With the same logic, we expect that $|f^{-k}(A) \cap B| \approx |A| \cdot |B|$ for any subinterval $B \subset [0, 1)$. Let us fix A and B , and denote by $\varepsilon_k = |f^{-k}(A) \cap B| - |A| \cdot |B|$ the error in this approximation: the smaller ε_k , the “better” shuffling we get after k steps. Then, $\frac{1}{n} \sum_{k=0}^{n-1} \varepsilon_k$ measures the *average* “quality” of shuffling after the first n steps. We say that our interval exchange procedure is *weakly mixing* if for every pair of measurable subsets $A, B \subset [0, 1)$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} \varepsilon_k = 0.$$

Intuitively, this means that the procedure works (on average) well, at least if we are ready to repeat it sufficiently many times.

What Procedures Are Weakly Mixing?

In summary, to program a robot to “shuffle” the interval $[0, 1)$, it suffices to construct a weakly mixing interval exchange procedure f (that is, construct intervals G_1, G_2, \dots, G_d and select a method, call it π , to permute them), and ask the robot to repeat it sufficiently many times. But what procedures are weakly mixing? For example, if our permutation π of G_1, G_2, \dots, G_d is reducible, then there is an $x \in (0, 1)$ such that points from $[0, x)$ never go to $[x, 1)$ and vice versa, therefore such a shuffling is obviously not “good”. If π is a rotation, we are in trouble for similar reasons. The following theorem of Artur Avila and Giovanni Forni [23] states that these are essentially the only “bad” cases!

Theorem 7.4 *Let π be an irreducible permutation which is not a rotation. Then for Lebesgue almost every $\lambda_1, \dots, \lambda_{d-1}$, the corresponding interval exchange procedure f is weakly mixing.*

Here, “Lebesgue almost every $\lambda_1, \dots, \lambda_{d-1}$ ” means that if we select $\lambda_1, \dots, \lambda_{d-1}$ at random, then the resulting interval exchange procedure will be weakly mixing with probability 1. For example, we can select $d = 3$, intervals $G_1 = [0, \lambda_1]$, $G_2 = [\lambda_1, \lambda_1 + \lambda_2]$, $G_3 = [\lambda_1 + \lambda_2, 1)$ with λ_1, λ_2 selected at random, a permutation $(G_1, G_2, G_3) \rightarrow (G_3, G_2, G_1)$, and be sure that if we repeat this procedure again and again, we will get a “permutation” of real numbers on $[0, 1)$ which is “well-shuffled”. Sometimes “well-shuffled” sets are easier to study, which makes Theorem 7.4 especially useful.

Reference

A. Avila and G. Forni, Weak mixing for interval exchange transformations and translation flows, *Annals of Mathematics* **165**-2, (2007), 637–664.

7.5 The Hopf Condition Over Arbitrary Fields

Which Integers Are the Sum of Two Squares?

Which integers can be represented as a sum of squares of two integers? For example, $0 = 0^2 + 0^2$, $1 = 1^2 + 0^2$, $2 = 1^2 + 1^2$, but there is no way to represent 3. Then $4 = 2^2 + 0^2$, $5 = 2^2 + 1^2$, but there is no such representation of 6 or 7, see Fig. 7.5b. The Greek mathematician Diophantus of Alexandria, who lived in the 3rd century AD, noticed an amazing fact: if integers x and y can be written as a sum of two squares, then their product xy can also be written in this way! Why so? Well, let $x = a^2 + b^2$, $y = c^2 + d^2$, then

$$xy = (a^2 + b^2)(c^2 + d^2) = (ac + bd)^2 + (ad - bc)^2. \quad (7.3)$$

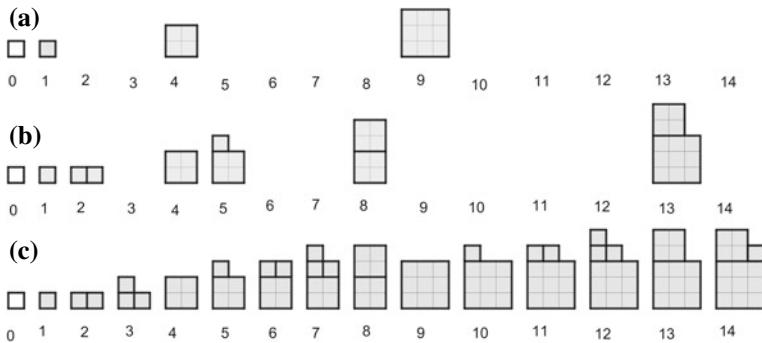


Fig. 7.5 Lists of **a** squares, **b** sums of two squares, **c** sums of four squares

Identity (7.3) can be proved by direct calculation – just perform the multiplication on the left and right-hand sides and check that the products are the same. This implies that xy is the sum of two squares whenever x and y are. For example, $5 = 2^2 + 1^2$, and $13 = 2^2 + 3^2$, hence, by (7.3),

$$65 = 5 \cdot 13 = (2^2 + 1^2)(2^2 + 3^2) = (2 \cdot 2 + 1 \cdot 3)^2 + (2 \cdot 3 - 1 \cdot 2)^2 = 7^2 + 4^2,$$

and actually this was the example mentioned by Diophantus.

How Many Squares are Needed to Represent Every Integer?

How many squares do we need to represent *every* non-negative integer? For 0, 1, 2, 4, 5 two squares suffice, $3 = 1^2 + 1^2 + 1^2$, $6 = 2^2 + 1^2 + 1^2$, but for 7 even three squares are not sufficient, so we need at least four. Then $7 = 2^2 + 1^2 + 1^2 + 1^2$, and indeed *every* positive integer is the sum of four squares, see Fig. 7.5c. This fact was known already to Diophantus, but the first rigorous proof was found by Lagrange in 1770. In fact, Lagrange proved this fact only for prime numbers, but this was sufficient because in 1748 Euler reported the formula

$$(a_1^2 + a_2^2 + a_3^2 + a_4^2)(b_1^2 + b_2^2 + b_3^2 + b_4^2) = \\ (a_1b_1 - a_2b_2 - a_3b_3 - a_4b_4)^2 + (a_1b_2 + a_2b_1 + a_3b_4 - a_4b_3)^2 + \\ (a_1b_3 - a_2b_4 + a_3b_1 + a_4b_2)^2 + (a_1b_4 + a_2b_3 - a_3b_2 + a_4b_1)^2. \quad (7.4)$$

This formula can be checked by direct calculation, and it implies that if x and y are sums of four squares, then so is their product xy . Because every integer is a product of primes, and primes are sums of four squares by Lagrange's theorem, this implies that *every* positive integer is the sum of four squares.

In Which Cases do Useful Identities Like (7.3) and (7.4) Exist?

Can we write a formula similar to (7.3) and (7.4) for products of three squares? Lagrange proved that

$$(a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2) = \\ (a_1b_1 + a_2b_2 + a_3b_3)^2 + (a_1b_2 - a_2b_1)^2 + (a_1b_3 - a_3b_1)^2 + (a_2b_3 - a_3b_2)^2, \quad (7.5)$$

but the right-hand side of this formula contains four squares instead of three, so (7.5) does not imply that “if x and y are sums of three squares, then so is xy ”. In fact, this is not the case, for example, 3 and 5 are the sums of three squares, but 15 is not. However, formula (7.5) turns out to be useful... in geometry, to study properties of the so-called “symmedian point” in a triangle.

Given that formulas like (7.3), (7.4), and (7.5) have so many useful applications, mathematicians became interested in values of (r, s, n) for which we can write

$$(a_1^2 + a_2^2 + \cdots + a_r^2)(b_1^2 + b_2^2 + \cdots + b_s^2) = z_1^2 + z_2^2 + \cdots + z_n^2, \quad (7.6)$$

where each z_i is a sum of the form

$$z_i = c_{i1}a_1b_1 + c_{i2}a_1b_2 + \cdots + c_{irs}a_rb_s = \sum_{j=1}^r \sum_{k=1}^s c_{ijk}a_jb_k$$

for some real coefficients c_{ijk} .

The Hopf Condition and Binomial Coefficients

Formulas (7.3), (7.4), and (7.5) show that (7.6) holds with $(r, s, n) = (2, 2, 2)$, $(r, s, n) = (4, 4, 4)$, and $(r, s, n) = (3, 3, 4)$, respectively. However, these are just the special cases. What we need is some general methods for determining for which triples (r, s, n) formula (7.6) is possible and for which it is not. One such general theorem was proved by Hopf [205] in 1941. It states that if (7.6) holds for some (r, s, n) then

(*) The numbers $\frac{n!}{i!(n-i)!}$ are even integers for all i such that $n - r < i < s$.

Here, $k!$ denotes the product of all integers from 1 to k (for example, $3! = 1 \cdot 2 \cdot 3 = 6$, $5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$, and so on), and the numbers $\frac{n!}{i!(n-i)!}$ are called *binomial coefficients*. These numbers are well-studied in combinatorics, because they denote the number of ways we can select i objects out of n . For example, if we need to select $i = 2$ people out of $n = 4$ (say, John, George, Emily, and Isabella), we can do this in 6 different ways, namely, John and George, John and Emily, John and Isabella, George and Emily, George and Isabella, or Emily and Isabella. And the formula gives us $\frac{4!}{2!(4-2)!} = \frac{1 \cdot 2 \cdot 3 \cdot 4}{1 \cdot 2 \cdot 1 \cdot 2} = 6$.

Ok, back to formula (7.6). Would such an identity be possible with, for example, $(r, s, n) = (5, 5, 5)$? If this were possible, then, by Hopf's condition (*), the numbers $\frac{5!}{i!(5-i)!}$ would be even integers for $5 - 5 < i < 5$. However, for $i = 1$, $\frac{5!}{1!4!} = 5$ is odd, a contradiction! By a similar argument we can exclude triples $(r, s, n) = (6, 6, 6)$, $(r, s, n) = (7, 7, 7)$, and, more generally, all triples in the form (n, n, n) except when n is a power of 2. In fact, there is a theorem that (7.6) is possible with $r = s = n$ only if $n = 1, 2, 4$, or 8.

Formula (7.6) Over Arbitrary Fields

So far, we have assumed that the variables (and coefficients c_{ijk}) in (7.6) are real numbers. In fact, we can write the same formula where the variables and coefficients are, for example, rational numbers, or complex numbers, or, more generally, elements of an arbitrary *field*. In short, a field is any set on which we can define the usual arithmetic operations (+, -, ·, /) in such a way that the usual properties hold. In particular, $x + y = y + x$, $xy = yx$, $x(y + z) = xy + xz$, etc. Also, there is a special element, called 0, such that $0 + x = x$ for all x , and another special element, called 1, such that $1 \cdot x = x$ for all x . In fact, the smallest possible field, called F_2 , contains just these two elements, 0 and 1, with addition and multiplication defined in the “normal” way except that $1 + 1 = 0$. This is called “addition modulo 2”. Similarly, F_3 is a field with three elements 0, 1, 2 and operations “modulo 3”, for example, $1 + 2 = 0$, $2 + 2 = 1$, $2 \cdot 2 = 1$, etc. See Sect. 1.7 for more details and more examples of fields.

In F_2 , $1 + 1 = 0$, while in F_3 , $1 + 1 + 1 = 0$. In general, the smallest number p such that $1 + 1 + \dots + 1$ (p times) = 0 in F is called the *characteristic* of the field F (if no such p exists we say that F has characteristic 0). For example, the field F_2 has characteristic $p = 2$, F_3 characteristic $p = 3$, while the field of all real (or rational) numbers has characteristic 0.

A Generalization of the Hopf Condition

In any field of characteristic 2, $x + x = x(1 + 1) = x \cdot 0 = 0$ for all x . Hence, $(x + y)^2 = x^2 + (xy + xy) + y^2 = x^2 + y^2$ for all x and y . This implies that $(a^2 + b^2)c^2 = (ac + bc)^2$, that is, formula (7.6) works for the triple $(r, s, n) = (2, 1, 1)$. In fact, by a similar argument, we can write such a formula for *all* triples (r, s, n) . In particular, nothing like Hopf's condition (*) can be true for such fields. However, it was an old conjecture that Hopf's condition remains true for all fields of characteristic $p \neq 2$. The following theorem of Dugger and Isaksen [122] confirms this conjecture.

Theorem 7.5 *If F is a field of characteristic not equal to 2, and a sums-of-squares formula (7.6) holds for some (r, s, n) , then the numbers $\frac{n!}{i!(n-i)!}$ are even integers for all i such that $n - r < i < s$.*

In general, the big classical question is: *given an arbitrary field F , and an arbitrary triple (r, s, n) , determine whether an identity (7.6) with these parameters exists, and if so, write it down.* With Theorem 7.5, we are now one big step closer to understanding this question. In particular, we can immediately deduce that (7.6) with, say, $(r, s, n) = (5, 5, 5)$ cannot hold in, for example, the field F_3 .

However, even with Theorem 7.5, there is still much to be done to achieve the complete answer to the classical question cited above. One problem is that the Hopf condition is necessary but not sufficient for the existence of a sums-of-squares formula. In particular, the triple $(r, s, n) = (13, 13, 16)$ satisfies the Hopf condition, but (7.6) cannot hold with these parameters in any field F of characteristic not equal to 2. This is because this triple does not satisfy another necessary condition, established by Dugger and Isaksen in another paper [121]: if (7.6) holds for some (r, s, n) , then $\frac{n!}{i!(n-i)!}$ is divisible by $2^{\lfloor \frac{s-1}{2} \rfloor - i + 1}$ for $n - r < i \leq \lfloor \frac{s-1}{2} \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer not exceeding x . For the triple $(r, s, n) = (13, 13, 16)$, it states that $\frac{16!}{i!(16-i)!}$ should be divisible by 2^{7-i} for $3 < i \leq 6$, but this is not true for $i = 4$, because $\frac{16!}{4!12!} = 1820$ is not divisible by $2^{7-4} = 8$.

Reference

D. Dugger and D. Isaksen, The Hopf condition for bilinear forms over arbitrary fields, *Annals of Mathematics* **165**-3, (2007), 943–964.

7.6 Any Real Polynomial Can be Approximated by a Hyperbolic Polynomial

Dynamical Systems and Their Fixed Points

The study of dynamical systems is an interesting area of mathematics which is the source of many problems and theorems which are easy to state but enormously hard to prove, see, for example, Sects. 2.7 and 5.4. In the simplest one-dimensional case, we are given a function $f : \mathbb{R} \rightarrow \mathbb{R}$, select an initial point x_0 , form a sequence by applying f to x_0 again and again

$$x_1 = f(x_0), \quad x_2 = f(x_1) = f(f(x_0)), \quad \dots, \quad x_n = f(x_{n-1}), \quad \dots, \quad (7.7)$$

and study the limiting behaviour of this sequence. This problem may be surprisingly difficult even if f is a very simple function, for example, a polynomial.

Of course, for some very simple polynomials the problem is easy. We start with an extremely simple case $f(x) = x^3$. If $x_0 = 0$, then $x_1 = x_0^3 = 0^3 = 0$, $x_2 = x_1^3 = 0^3 = 0$, and so on, $x_n = 0$ for all n . In general, if $f(x_0) = x_0$, then the whole sequence (7.7) looks like $x_0, x_0, \dots, x_0, \dots$. In this case, we say that x_0 is a *fixed point* of f . Our function $f(x) = x^3$ has two more fixed points, -1 and 1 . Indeed, in this case,

the equation $f(x_0) = x_0$ reduces to $x_0^3 = x_0$, or $x_0^3 - x_0 = 0$, or $x_0(x_0^2 - 1) = 0$, or $x_0(x_0 - 1)(x_0 + 1) = 0$, thus x_0 is equal to 0, -1, or 1.

Attracting and Non-attracting Fixed Points

What if we start sequence (7.7) from point x_0 which is *not* a fixed point? For example, let $x_0 = \frac{1}{2}$, then

$$x_1 = x_0^3 = \frac{1}{8}, \quad x_2 = x_1^3 = \frac{1}{512}, \quad \dots, \quad x_n = \frac{1}{2^{3^n}}, \quad \dots$$

and we can see that the sequence quickly converges to the fixed point 0. In fact, $x_n = (x_0)^{3^n}$, and therefore $x_n \rightarrow 0$ for any $x_0 \in (-1, 1)$, and $|x_n| \rightarrow \infty$ for any $x_0 > 1$ or $x_0 < -1$. In particular, x_n converges to 1 only if $x_0 = 1$, and to -1 only if $x_0 = -1$.

In general, a fixed point x^* is called *attracting* if there exists an $\varepsilon > 0$ such that sequence (7.7) converges to x^* for any x_0 such that $|x_0 - x^*| < \varepsilon$. So, in our example, 0 is an attracting fixed point with $\varepsilon = 1$, while the fixed points -1 and 1 are not attracting.

A Derivative-Based Test for “Attractiveness”

In fact, there is a deep *reason* which makes a fixed point attracting or not. Intuitively, a fixed point x^* is attracting if the distance to it decreases after each iteration. That is, $|x_{n+1} - x^*| < |x_n - x^*|$. But $x_{n+1} = f(x_n)$ and $x^* = f(x^*)$, hence this inequality can be rewritten as $|f(x_n) - f(x^*)| < |x_n - x^*|$. Or, denoting $x_n - x^*$ as ε , we get

$$\left| \frac{f(x^* + \varepsilon) - f(x^*)}{\varepsilon} \right| < 1. \quad (7.8)$$

If $\varepsilon \rightarrow 0$, the left-hand side in (7.8), is, by definition, the (absolute value of) the *derivative* of the function f at the point x^* , usually denoted by $f'(x^*)$. Condition (7.8) is then guaranteed to hold for all small ε if $|f'(x^*)| < 1$. In this case, we know that x^* is attracting. For similar reasons, if $|f'(x^*)| > 1$, we know that x^* is *not* attracting. The only tricky case is $|f'(x^*)| = 1$, in which case we cannot decide if a fixed point is attracting or not based only on the information about the derivative.

There is a simple formula for calculating the derivative of a polynomial: $(x^n)' = nx^{n-1}$ for any n . For $n = 3$, we get $(x^3)' = 3x^2$. Hence, for $f(x) = x^3$ we have $f'(0) = 3 \cdot 0^2 = 0 < 1$, but $f'(-1) = f'(1) = 3 > 1$. That is why the fixed point 0 is attracting, while the points -1 and 1 are not.

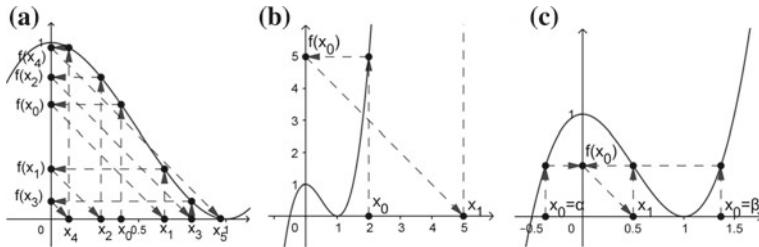


Fig. 7.6 Examples of dynamics of $f(x) = 2x^3 - 3x^2 + 1$ for various x_0

Attracting Periodic Points

With this machinery at hand, let us try to analyse the dynamics of more complicated polynomials, for example, $f(x) = 2x^3 - 3x^2 + 1$. To find its fixed points, we should solve the equation $f(x) = x$, that is, $2x^3 - 3x^2 + 1 = x$. One of the solutions is $x = 1/2$ (indeed, $f(1/2) = 2(1/2)^3 - 3(1/2)^2 + 1 = 2/8 - 3/4 + 1 = 1/2$). Now, let us check if the fixed point $1/2$ is attracting. The derivative $f'(x) = 2(x^3)' - 3(x^2)' = 6x^2 - 6x$. Hence, $f'(1/2) = 6(1/2)^2 - 6(1/2) = -3/2$, and $|f'(1/2)| = 3/2 > 1$. Thus, the fixed point $x = 1/2$ is not attracting. There are two more fixed points for f , but in the same way we can check that they are also not attracting! So, where does the sequence (7.7) converge to in this case?

Let us do an experiment starting with, say, $x_0 = 0.4$. Then

$$x_1 = f(x_0) = 0.648, \quad x_2 \approx 0.28, \quad x_3 \approx 0.80, \quad x_4 \approx 0.10, \quad x_5 \approx 0.97, \dots,$$

and then $x_n \approx 1$ for odd n , while $x_n \approx 0$ for even n , see Fig. 7.6a. Hence, the sequence x_n does not converge to any limit, but rather oscillates between 0 and 1. In fact, if we start with $x_0 = 0$, then $x_1 = 2 \cdot 0^3 - 3 \cdot 0^2 + 1 = 1$, $x_2 = 2 \cdot 1^3 - 3 \cdot 1^2 + 1 = 0$, $x_3 = 1$, $x_4 = 0$, and so on. In general, if $x_k = x_0$ for some k , then x_0 is called a *periodic point* of f with period k . Let us denote the function $f(f(x))$ by $f^2(x)$, $f(f(f(x)))$ by $f^3(x)$, and so on, so that $x_n = f^n(x)$ in (7.7). A periodic point x^* is called *attracting* if there exists an $\varepsilon > 0$ such that $\lim_{n \rightarrow \infty} |f^n(x_0) - f^n(x^*)| = 0$ whenever $|x_0 - x^*| < \varepsilon$.

Hyperbolic Points, Sets, and Polynomials

How do we check if a periodic point x^* with period k is attracting? In fact, $x_k = x_0$ implies that $f^k(x_0) = x_0$, that is, x_0 is a fixed point for a function $g(x) = f^k(x)$. Thus, we can guarantee that it is attracting if $|g'(x^*)| < 1$, and is not attracting if $|g'(x^*)| > 1$. Hence, only in the tricky case $|g'(x^*)| = 1$ can we not decide whether it is attracting or not. A periodic point x^* which avoids this tricky situation, that is, such that $|g'(x^*)| \neq 1$, is called *hyperbolic*.

For every f , including $f(x) = 2x^3 - 3x^2 + 1$, and for every x_0 , one of three cases is possible.

- (i) The sequence (7.7) may converge to a hyperbolic attracting periodic point, like for $x_0 = 0.4$ in Fig. 7.6a;
- (ii) $|x_n|$ in (7.7) may converge to infinity, for example, for $x_0 = 2, x_1 = 2 \cdot 2^3 - 3 \cdot 2^2 + 1 = 5, x_2 = 176, x_3 = 10810625$, and so on, see Fig. 7.6b;
- (iii) x_0 belongs to some set S (let us call it an “erratic set”) where neither (i) nor (ii) happen. For example, $1/2 \in S$. Also, equation $f(x) = 1/2$ has, besides $1/2$, two other solutions, say α and β . Then for $x_0 = \alpha$ (or $x_0 = \beta$) we have $x_1 = f(\alpha) = 1/2, x_2 = f(1/2) = 1/2, x_3 = f(1/2) = 1/2$, and so on, hence $\alpha \in S$ and $\beta \in S$, see Fig. 7.6c. Then the solutions to $f(x) = \alpha$ (or $f(x) = \beta$) also belong to S , and so on.

A set S is called *hyperbolic* if there exist constants $C > 0$ and $\lambda > 1$ such that $|(f^n)'(x)| > C\lambda^n, \forall n, \forall x \in S$. A polynomial f is called *hyperbolic* if the corresponding “erratic set” S in (iii) is hyperbolic. If f is hyperbolic, then, for most x_0 , either (i) or (ii) happens, hence the sequence (7.7) is easy to analyse. For example, $f(x) = x^3$ and $f(x) = 2x^3 - 3x^2 + 1$ are hyperbolic polynomials.

Hyperbolic Polynomials are “Dense”

Not all polynomials are hyperbolic. For example, for $f(x) = x^3 - 3x$ and $x_0 = 0.5$, the first terms of sequence (7.7) are

$$x_1 = -1.375, \quad x_2 \approx 1.53, \quad x_3 \approx -1.03, \quad x_4 \approx 1.99, \quad x_5 \approx 1.98, \quad x_6 \approx 1.82, \dots,$$

and so on, the sequence converges neither to a number nor to a cycle. And the same is true for almost every x_0 chosen from $[-2, 2]$. In other words, in this case the “erratic set” S (almost) coincides with the whole interval $[-2, 2]$. In particular, S is not a hyperbolic set, and, by definition, f is not a hyperbolic polynomial.

The following theorem, proved in [231], states that, for any polynomial f which is not hyperbolic, we can find some hyperbolic polynomial h which is very “close” to f .

Theorem 7.6 *Any real polynomial f can be approximated by hyperbolic real polynomials of the same degree. That is, if $f(x) = a_dx^d + a_{d-1}x^{d-1} + \dots + a_1x + a_0$, then, for any $\varepsilon > 0$, there exists a hyperbolic polynomial $h(x) = b_dx^d + b_{d-1}x^{d-1} + \dots + b_1x + b_0$ such that $|a_i - b_i| < \varepsilon, i = 0, 1, \dots, d$.*

In 2000, Smale [348] published a list of problems which he thought were the most important open problems in mathematics for the 21st century. The second part of Smale’s eleventh problem asks if every “sufficiently smooth” function f can be approximated by hyperbolic functions. Theorem 7.6 immediately resolves this problem! Indeed, it is known that any such f can be approximated by polynomials,

which, by Theorem 7.6, can in turn be approximated by hyperbolic polynomials. Done!

Reference

O. Kozlovski, W. Shen, and S. van Strien, Density of hyperbolicity in dimension one, *Annals of Mathematics* **166**-1, (2007), 145–182.

7.7 The Schinzel–Zassenhaus Conjecture for Polynomials with Odd Coefficients

Reducible and Irreducible Polynomials

As you know, an integer m is called a *divisor* of an integer n if $n = mk$ for some integer k . An integer $n > 1$ which can be written as $n = mk$ with integers $m > 1$ and $k > 1$ is called *composite*, and otherwise it is called *prime*. Every positive integer can be written in a unique way as a product of primes. For example, $12 = 4 \cdot 3$, hence 3 and 4 are divisors of 12. 3 cannot be factorized further, it is prime. 4 can be written as $2 \cdot 2$, and the prime factorization of 12 is $2 \cdot 2 \cdot 3$.

The same definitions can be introduced for polynomials with integer coefficients, that is, functions $P(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$, where $a_n, a_{n-1}, \dots, a_1, a_0$ are integers. A polynomial $Q(x)$ is a divisor of $P(x)$ if $P(x) = Q(x) \cdot R(x)$, where $R(x)$ is another polynomial with integer coefficients. For example, $x^2 - 1 = (x - 1)(x + 1)$, hence $x - 1$ and $x + 1$ are divisors of $x^2 - 1$. A polynomial $P(x)$ which can be written as a product $Q(x) \cdot R(x)$ of two non-constant polynomials with integer coefficients is called *reducible*, and otherwise it is called *irreducible*. For example, the polynomial $x^2 - 1$ is reducible (because $x^2 - 1 = (x - 1)(x + 1)$), while $x^2 + 1$ is irreducible.

Factorization of $x^n - 1$ and Cyclotomic Polynomials

Every polynomial can be written as a product of irreducible polynomials. Let us practice with polynomials of the form $x^n - 1$.

- (i) $n = 1$: $x - 1$ is irreducible;
- (ii) $n = 2$: $x^2 - 1 = (x - 1)(x + 1)$;
- (iii) $n = 3$: $x^3 - 1 = (x - 1)(x^2 + x + 1)$;
- (iv) $n = 4$: $x^4 - 1 = (x^2 - 1)(x^2 + 1) = (x - 1)(x + 1)(x^2 + 1)$;
- (v) $n = 5$: $x^5 - 1 = (x - 1)(x^4 + x^3 + x^2 + x + 1)$;
- (vi) $n = 6$: $x^6 - 1 = (x^3 - 1)(x^3 + 1) = (x - 1)(x^2 + x + 1)(x + 1)(x^2 - x + 1)$;
- (vii) $n = 7$: $x^7 - 1 = (x - 1)(x^6 + x^5 + x^4 + x^3 + x^2 + x + 1)$;

$$(viii) \ n = 8: \ x^8 - 1 = (x^4 - 1)(x^4 + 1) = (x^2 - 1)(x^2 + 1)(x^4 + 1) = (x - 1)(x + 1)(x^2 + 1)(x^4 + 1),$$

and so on. Note that, for example, term $x^4 + 1$ in the last product can be further factorized as $x^4 + 1 = (x^2 + \sqrt{2}x + 1)(x^2 - \sqrt{2}x + 1)$, but this “does not count”, because we allow only factorization by polynomials with *integer* coefficients.

Irreducible polynomials which are divisors of $x^n - 1$ for some n are called *cyclotomic* polynomials. Hence, the cyclotomic polynomials are $x - 1, x + 1, x^2 + x + 1, x^2 + 1, x^4 + x^3 + x^2 + x + 1, x^2 - x + 1, x^6 + x^5 + x^4 + x^3 + x^2 + x + 1, x^4 + 1$, and so on. Interestingly, the factorization for each $x^n - 1$ produces exactly one new cyclotomic polynomial, which does not appear in any factorization of $x^k - 1$, $k < n$.

A polynomial $P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ of degree n is called *monic* if $a_n = 1$. For example, the polynomials $x + 2$ and $x^2 - 3x + 4$ are monic, while $2x^2 - 3x + 1$ are not. All cyclotomic polynomials are monic.

Roots of Cyclotomic Polynomials

A *root* of a polynomial $P(x)$ is a value of x such that $P(x) = 0$. Every linear polynomial $P(x) = ax + b, a \neq 0$, has exactly one root $x = -b/a$. The quadratic polynomial $P(x) = ax^2 + bx + c$ with $b^2 - 4ac > 0$ has two roots

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \quad (7.9)$$

However, if $b^2 - 4ac < 0$, the polynomial $ax^2 + bx + c$ has no real roots. For example, this is the case if $b = 0, a = c = 1, 0^2 - 4 \cdot 1 \cdot 1 < 0$, and the polynomial $x^2 + 1$ has no real roots. For this reason, mathematicians introduced *complex* numbers, that is, numbers of the form $x + yi$, where x, y are real numbers, and i is the imaginary “square root of minus one”, that is, a number such that $i^2 = -1$. Using complex numbers, we can extract the square root of *any* real number (for example, $\sqrt{-9} = \sqrt{9 \cdot (-1)} = \sqrt{9} \cdot \sqrt{-1} = 3i$), and thus we can use formula (7.9) to find two complex roots x_1 and x_2 of *any* quadratic polynomial (we can have $x_1 = x_2$ if $b^2 - 4ac = 0$). For example, let us find the roots of the cyclotomic quadratic polynomials listed above.

- (a) For $x^2 + x + 1$, the roots are $x_{1,2} = \frac{-1 \pm \sqrt{1^2 - 4 \cdot 1 \cdot 1}}{2 \cdot 1} = -0.5 \pm (\sqrt{3}/2)i$;
- (b) For $x^2 + 1$, $x_{1,2} = \frac{0 \pm \sqrt{0^2 - 4 \cdot 1 \cdot 1}}{2 \cdot 1} = \pm i$;
- (c) For $x^2 - x + 1$, $x_{1,2} = \frac{1 \pm \sqrt{(-1)^2 - 4 \cdot 1 \cdot 1}}{2 \cdot 1} = 0.5 \pm (\sqrt{3}/2)i$.

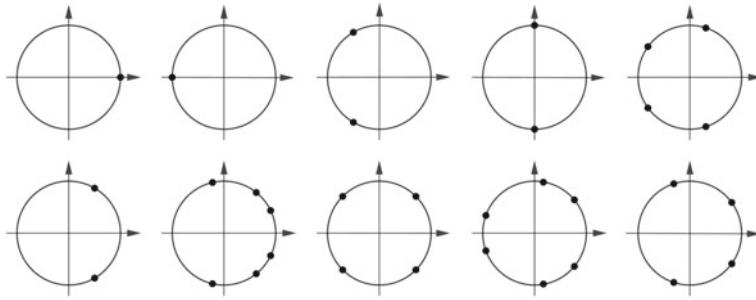


Fig. 7.7 Roots of the first 8 cyclotomic polynomials

The Absolute Value of the Roots

The *absolute value* $|z|$ of a complex number $z = x + yi$ is, by definition, $|z| = \sqrt{x^2 + y^2}$. For example, if $y = 0$, then $z = x$ is a real number, and $|z| = \sqrt{x^2 + 0^2} = |x|$ is just the “usual” absolute value of the real number x . For complex numbers, the absolute value satisfies some of the “usual” properties, for example, $|z| \geq 0$ for every z , and $|z_1 \cdot z_2| = |z_1| \cdot |z_2|$ for any two complex numbers z_1 and z_2 . Applying the latter property n times, we can conclude that $|z^n| = |z|^n$. For the roots of the cyclotomic polynomials calculated in (a), (b), (c), the absolute values are:

- $| -0.5 \pm (\sqrt{3}/2)i | = \sqrt{(-0.5)^2 + (\pm\sqrt{3}/2)^2} = \sqrt{1/4 + 3/4} = 1,$
- $| \pm i | = | 0 \pm 1 \cdot i | = \sqrt{0^2 + (\pm 1)^2} = 1,$
- $| 0.5 \pm (\sqrt{3}/2)i | = \sqrt{0.5^2 + (\pm\sqrt{3}/2)^2} = 1.$

The roots of the first 8 cyclotomic polynomials are depicted in Fig. 7.7, and, in all cases, the absolute value of each root is 1. This is not a coincidence! In fact, all roots of all cyclotomic polynomials always have absolute value 1. Indeed, if $P(x)$ is any cyclotomic polynomial, then, by definition, $P(x)$ is a divisor of $x^n - 1$ for some n , that is, $x^n - 1 = P(x)Q(x)$. Now, if z is any root of $P(x)$, then $P(z) = 0$, hence $z^n - 1 = P(z)Q(z) = 0 \cdot Q(z) = 0$, thus $z^n = 1$. This implies that $|z^n| = |1| = 1$. But $|z^n| = |z|^n$, hence $|z|^n = 1$, and $|z| = \sqrt[n]{1} = 1$.

Of course, if the polynomial $P(x)$ is not cyclotomic but can be written as a product of cyclotomic polynomials

$$P(x) = Q_1(x) \cdot Q_2(x) \dots Q_k(x), \quad Q_i(x) \text{ are cyclotomic, } i = 1, 2, \dots, k, \tag{7.10}$$

then all roots of $P(x)$ are roots of $Q_i(x)$ and therefore also have absolute value 1. For example, the polynomial $x^3 + 1$ can be written as a product $(x + 1)(x^2 - x + 1)$ of cyclotomic polynomials, and its roots $x_1 = -1$, $x_{2,3} = 0.5 \pm (\sqrt{3}/2)i$ all have absolute value 1.

A Conjecture of Schinzel and Zassenhaus

In 1857, Kronecker proved that if a monic polynomial $P(x)$ with integer coefficients cannot be represented in the form (7.10), then it has a root z with absolute value strictly greater than 1. In 1965, Schinzel and Zassenhaus [334] conjectured that, for every such polynomial of degree n , we can in fact find a root with absolute value greater than $1 + \frac{C}{n}$ for some absolute constant $C > 0$.

For example, any root z of the polynomial $x^n - 2 = 0$ (which is monic, with integer coefficients, and not in the form (7.10)) satisfies the equation $z^n = 2$, hence $|z^n| = |z|^n = 2$, and $|z| = \sqrt[n]{2}$. However, one can prove that $\sqrt[n]{2} > 1 + \frac{\ln 2}{n}$ for all n , hence, in this case, the conjecture holds with $C = \ln 2 \approx 0.69$.

For about 40 years, there was little progress towards a resolution of this conjecture. The following theorem of Borwein, Dobrowolski, and Mossinghoff [69] resolves it at least in the case when all coefficients of $P(x)$ are odd.

Theorem 7.7 *Let $P(x) = x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$ be monic polynomial such that all coefficients a_0, a_1, \dots, a_{n-1} are odd integers. If $P(x)$ cannot be represented as a product of cyclotomic polynomials, as in (7.10), then it has a root z with $|z| > 1 + \frac{\ln 3}{2(n+1)}$.*

Because $1 + \frac{\ln 3}{2(n+1)} \geq 1 + \frac{(\ln 3)/4}{n}$, Theorem 7.7 implies the Schinzel–Zassenhaus conjecture for polynomials with odd coefficients. The condition of “odd coefficients” can be written as $a_i = b_i \cdot 2 + 1$ for some integers b_i , $i = 0, \dots, n-1$. In fact, Borwein et al. also proved the Schinzel–Zassenhaus conjecture for polynomials with coefficients in the form $a_i = b_i \cdot m + 1$, where $m \geq 2$ is any integer. For example, the polynomial $x^2 + 4x + 10$ satisfies this condition for $m = 3$, and therefore is also covered by their theorem.

A Partial Answer to Lehmer’s Question

As often happens, techniques developed to prove one result turn out to have applications to other related questions. Mahler’s measure $M(P)$ of a polynomial P is the product of the absolute values of those roots of P whose absolute value is at least 1, multiplied by the absolute value of the leading coefficient. For example, $P(x) = 4x^3 - 9x$ has roots $-3/2, 0, 3/2$, with absolute values $3/2, 0$, and $3/2$, respectively. Hence, $M(P) = (3/2) \cdot (3/2) \cdot 4 = 9$. In 1933, Lehmer found the polynomial

$$P^*(x) = x^{10} + x^9 - x^7 - x^6 - x^5 - x^4 - x^3 + x + 1$$

with $M(P^*) = 1.176280\dots$, but could not find a polynomial P with integer coefficients such that $1 < M(P) < M(P^*)$. Motivated by this, he asked if such a polynomial exists, and, more generally, if we can find P satisfying $1 < M(P) < 1 + \varepsilon$ for every $\varepsilon > 0$. In the same paper [69], the authors answered Lehmer’s question negatively for polynomials with odd coefficients.

Reference

P. Borwein, E. Dobrowolski, and M. Mossinghoff, Lehmer’s problem for polynomials with odd coefficients, *Annals of Mathematics* **166**-2, (2007), 347–366.

7.8 Diophantine Approximation of Points on Smooth Planar Curves

Approximating the Area of the Unit Circle

It has been known for a long time that the area of the circle with radius 1 meter cannot be “measured” exactly. This area, which mathematicians denote by π , is equal to 3 square metres, and 1415 square centimetres, and 9265 square millimetres, and so on, whatever small unit of measurements we choose, it will be a such units plus a remainder, and this remainder will never be 0. This is because π is an example of an *irrational* number, that is, a number not representable in the form $\frac{a}{b}$ for integers a and b .

Of course, any irrational number can be approximated by a rational number (that is, by $\frac{a}{b}$) arbitrary well. For example, $\pi \approx \frac{31}{10}$ with error less than $\frac{1}{20}$, $\pi \approx \frac{3142}{1000}$ with error less than $\frac{1}{2000}$, and, in general, for any b we can select a such that $|\pi - \frac{a}{b}| < \frac{1}{2b}$. Indeed, this inequality is equivalent to $|b\pi - a| < \frac{1}{2}$, so it is sufficient to choose a to be the nearest integer to $b\pi$.

Of course, it is easy to select b such that the distance from $b\pi$ to the nearest integer is much smaller than $\frac{1}{2}$. Let us write down $b\pi$ for $b = 1, 2, 3, \dots$

$$\pi \approx 3.14, \quad 2\pi \approx 6.28, \quad 3\pi \approx 9.42, \quad 4\pi \approx 12.57, \quad 5\pi \approx 15.71, \quad 6\pi \approx 18.85, \quad 7\pi \approx 21.99, \dots \quad (7.11)$$

We can see that $b\pi$ is especially close to an integer for $b = 7$. In this case, $|7\pi - 22|$ is much less than $\frac{1}{2}$, leading to a very good approximation $\pi \approx \frac{22}{7}$.

The Quality of Approximation: From the Golden Ratio to Liouville’s Constant

In fact, we can find b with $|b\pi - a|$ as small as we want, and this is true not only for π . The famous Dirichlet approximation theorem states that for any irrational number α there are infinitely many values of a and b such that $|b\alpha - a| < \frac{1}{b}$, which implies that $|\alpha - \frac{a}{b}| < \frac{1}{b^2}$. In 1903, Borel improved this to $|b\alpha - a| < \frac{1}{\sqrt{5}b}$, or $|\alpha - \frac{a}{b}| < \frac{1}{\sqrt{5}b^2}$. A natural question is if this estimate can be further improved, yielding an even better approximation. For example, can we find infinitely many a and b such that $|b\alpha - a| < \frac{0.001}{b}$, or $|b\alpha - a| < \frac{1}{b^2}$, or maybe even $|b\alpha - a| < \frac{1}{2^b}$? In general, for which decreasing functions ψ do we have

$$|b\alpha - a| < \psi(b) \quad \text{for infinitely many pairs } (a, b)? \quad (7.12)$$

It turns out that there are some special irrational numbers, for example the “golden ratio” $\alpha = \frac{1+\sqrt{5}}{2}$, for which Borel’s estimate cannot be improved at all. On the other hand, it is easy to construct other special irrational numbers which can be approximated by rationals extremely well. For example, Liouville’s constant is the number

$$L = 0.11000100000000000000001000\dots$$

with digits 0 and 1 only, such that the k -th digit is 1 if and only if $k = 1 \cdot 2 \cdots (n-1) \cdot n$ for some n . For $n = 1, 2, 3, 4, 5$ this gives $k = 1, 2, 6, 24, 120$, hence the next 1 in L is the 120-th digit. Then $L \approx \frac{11}{100}$ with error about 10^{-6} , while $L \approx \frac{110001}{1000000}$ with error about 10^{-24} , and so on. In each case we have an error *much* smaller than the denominator b . In particular, (7.12) holds infinitely often even with $\psi(b) = 1/b^{100}$, or in fact for $\psi(b) = 1/b^N$ for any fixed N , no matter how large.

The Quality of Approximation of “Most” Irrational Numbers

However, at which accuracy we can approximate “most” irrational numbers, not just some special ones like Liouville’s constant? In 1925, Khinchin [222] gave an elegant and general answer to this question. His theorem states that if α is chosen from any interval $[a, b]$ ($a < b$) uniformly at random, then, with probability 1, (7.12) holds if and only if $\sum_{b=1}^{\infty} \psi(b) = \infty$. Here, by $\sum_{b=1}^{\infty} \psi(b)$ we mean $\lim_{n \rightarrow \infty} \sum_{b=1}^n \psi(b)$. For example, if $\psi(b) = \frac{1}{2}$ for all b , then $\sum_{b=1}^n \psi(b) = n/2$, and $\sum_{b=1}^{\infty} \psi(b) = \lim_{n \rightarrow \infty} n/2 = \infty$, in agreement with our (trivial) observation above that (7.12) holds for $\phi(b) = \frac{1}{2}$. A slightly less trivial computation shows that $\sum_{b=1}^{\infty} \psi(b) = \infty$ for $\psi(b) = \frac{0.001}{b}$, and in fact for $\psi(b) = \frac{\varepsilon}{b}$ for any $\varepsilon > 0$, hence (7.12) holds for these $\psi(b)$ as well. On the other hand, if $\psi(b) = 2^{-b}$, then $\sum_{b=1}^n \psi(b) = \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^n} = 1 - \frac{1}{2^n}$, and $\sum_{b=1}^{\infty} \psi(b) = \lim_{n \rightarrow \infty} (1 - \frac{1}{2^n}) = 1 < \infty$. With a bit more effort, one can show that $\sum_{b=1}^{\infty} \psi(b) < \infty$ for $\psi(b) = 1/b^{100}$, $\psi(b) = 1/b^2$, and in fact for $\psi(b) = 1/b^{1+\varepsilon}$ for any $\varepsilon > 0$. Hence, these $\psi(b)$ decay “too fast” for (7.12) to hold.

Dirichlet and Khintchine’s Theorems for Simultaneous Approximation

Now, can we approximate *two* different irrational numbers, say π and π^5 , by fractions $\frac{a_1}{b}$ and $\frac{a_2}{b}$ with the same denominator? In other words, can we find b such that $b\pi$ and

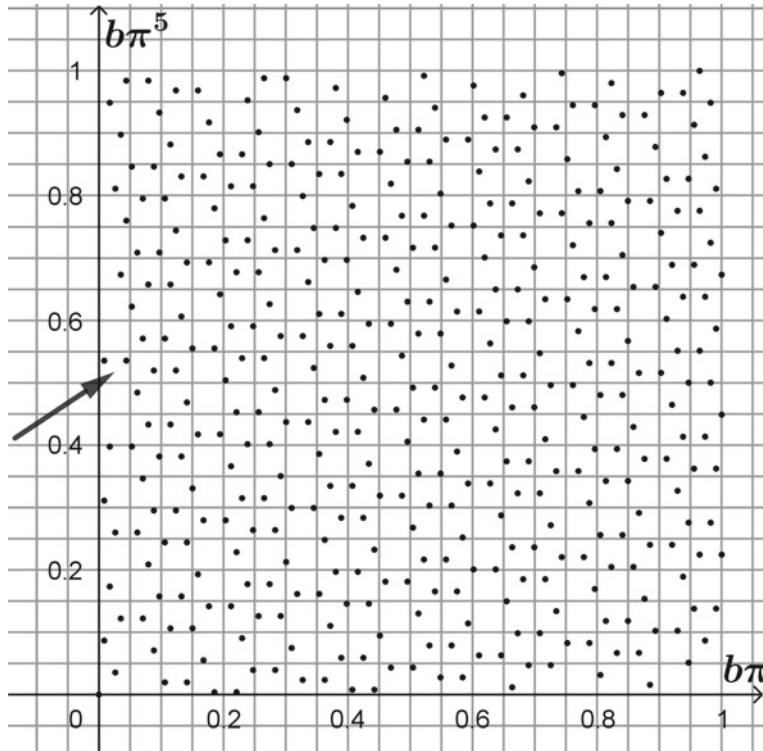


Fig. 7.8 The two-dimensional Dirichlet approximation theorem for $(\alpha, \beta) = (\pi, \pi^5)$

$b\pi^5$ are simultaneously close to an integer, e.g., find b such that $|b\pi - a_1| < 0.05$ and $|b\pi^5 - a_2| < 0.05$?

In fact, we can, and this is easy to see. For $b = 0, 1, 2, \dots, 400$, mark on the coordinate plane the point with coordinates $(\{b\pi\}, \{b\pi^5\})$, where $\{x\}$ denotes the fractional part¹ of x . Because $0 \leq \{x\} < 1$ for all x , all 401 marked points lie in the unit square, see Fig. 7.8. If we divide this unit square into $20^2 = 400$ sub-squares with side length $\frac{1}{20} = 0.05$ each, then, by the Pigeonhole principle, there are two points in the same sub-square. If these points are $(\{b_1\pi\}, \{b_1\pi^5\})$ and $(\{b_2\pi\}, \{b_2\pi^5\})$, then $|\{b_1\pi\} - \{b_2\pi\}| < 0.05$, and $|\{b_1\pi^5\} - \{b_2\pi^5\}| < 0.05$. But this means that the numbers $(|b_1 - b_2|)\pi$ and $(|b_1 - b_2|)\pi^5$ are at distance at most 0.05 from some integers a_1 and a_2 , respectively.

The same argument shows that for any pair of numbers (α, β) there are infinitely many triples of integers (b, a_1, a_2) such that

¹That is, $\{x\} = x - [x]$, where $[x]$ is the largest integer not exceeding x . For example, $[\pi] = 3$, and $\{\pi\} = 0.14159\dots$

$$\max(|b\alpha - a_1|, |b\beta - a_2|) < \frac{1}{\sqrt{b}}.$$

This is known as two-dimensional version of Dirichlet's approximation theorem. Similar to (7.12), we can ask if this estimate can be improved, and if so, by how much, that is, for which decreasing functions ψ do we have

$$\max(|b\alpha - a_1|, |b\beta - a_2|) < \psi(b) \quad \text{for infinitely many triples } (b, a_1, a_2)? \quad (7.13)$$

A two-dimensional version of Khinchin's theorem gives the answer: if (α, β) is chosen from any rectangle $[x, y] \times [z, w]$ ($x < y, z < w$) uniformly at random, then, with probability 1, (7.13) holds if and only if

$$\sum_{b=1}^{\infty} \psi(b)^2 = \infty. \quad (7.14)$$

Simultaneous Approximation of “Related” Irrational Numbers

However, what if (α, β) is *not* a random pair of numbers, but are *related* to each other in some way? In our example with (π, π^5) , the second number is the fifth power of the first one. In general, what if $\beta = \alpha^5$, or $\beta = e^\alpha$, or, more generally, $\beta = f(\alpha)$ for some function f ? The following theorem of Beresnevich, Dickinson, and Velani [42] addresses this question for a broad class of functions f . If some property holds for a random point with probability 1, we will say that it holds for *almost all* points.

Theorem 7.8 *Let $\psi : [0, \infty) \rightarrow [0, \infty)$ be a decreasing function satisfying (7.14). Let f be a three times continuously differentiable function on some interval (a, b) , such that $f''(x) \neq 0$ for almost all $x \in (a, b)$. Then, for almost all $x \in (a, b)$, (7.13) holds for $(\alpha, \beta) = (x, f(x))$.*

In case you do not know the term “continuously differentiable” and the notation $f''(x)$, see e.g. Sect. 3.1 for the relevant definitions. In short, most “nice” functions like $f(x) = x^2$, $f(x) = x^5$, or $f(x) = e^x$, satisfy the conditions of the theorem. In particular, we now know that (7.13) holds for almost all α , if, for example, $\beta = \alpha^5$, and $\psi(b) = \frac{\varepsilon}{\sqrt{b}}$ for any $\varepsilon > 0$.

Reference

V. Beresnevich, D. Dickinson, and S. Velani, Diophantine approximation on planar curves and the distribution of rational points, *Annals of Mathematics* **166**-2, (2007), 367–426.

7.9 The Hasse Principle for Systems of Two Diagonal Cubic Equations

The Search for a Non-Trivial Integer Solution

Finding integer solutions to polynomial equations in several variables is a notoriously difficult problem. There is even a formal theorem proving that there is no general method (algorithm) which would work for *all* such equations. However, there are many powerful methods which work at least for some equations.

For illustration, let us start with a relatively easy question: for what values of the parameters a and b does the equation

$$ax^2 + by^2 = 0 \quad (7.15)$$

have a solution different from $x = y = 0$? The solution $x = y = 0$ is called *trivial*, so we are looking for non-trivial solutions.

No Real Solutions Implies No Integer Solutions

First, note that if (x_0, y_0) is a non-trivial solution, which has common divisor $d > 1$, then $x_1 = x_0/d$, $y_1 = y_0/d$ is a solution as well. Indeed, $ax_1^2 + by_1^2 = a(x_0/d)^2 + b(y_0/d)^2 = (ax_0^2 + by_0^2)/d^2 = 0/d^2 = 0$, where the third equality follows from the fact that (x_0, y_0) is a solution. If (x_1, y_1) has another common divisor $d_1 > 1$, we can divide by it to find one more solution, and, ultimately, we will always find a non-trivial solution (x, y) such that x and y have no common divisor greater than 1 (such numbers are called relatively prime).

Next, let us think about possible values of parameters a and b . Of course, if $a = 0$, then, for example, $x = 1, y = 0$ is a non-trivial solution. Similarly, if $b = 0$, then $x = 0, y = 1$ works. So, let us assume that $a \neq 0, b \neq 0$.

If a and b are both positive, for example, $a = b = 1$, then $ax^2 + by^2 \geq 0$, and in fact $ax^2 + by^2 > 0$ unless $x = y = 0$. Hence, in this case, the equation has no non-trivial *real* solution. Of course, this implies that there are no non-trivial integer solutions as well. Similarly, if a and b are both negative, then $ax^2 + by^2 < 0$ unless $x = y = 0$, so there are no non-trivial solutions in this case as well.

Odd-Even Analysis

The remaining case is when a and b have different sign, for example, $a = 1, b = -2$. In this case, the equation $x^2 - 2y^2 = 0$ has many non-trivial real solutions (for example, $x = \sqrt{2}, y = 1$), but it has no integer solutions, see Fig. 7.9b. To prove this, assume that the equation has a relatively prime solution (x, y) , and let us analyse whether x and y are even or odd. Because $x^2 = 2y^2$, x^2 is even, and x should be even

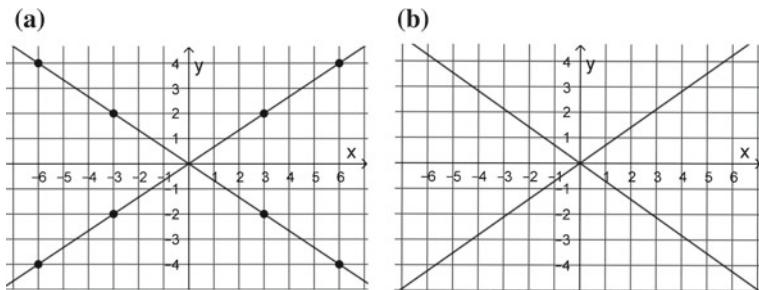


Fig. 7.9 Real and integer solutions of **a** $4x^2 - 9y^2 = 0$ and **b** $x^2 - 2y^2 = 0$

as well. Write $x = 2z$ for integer z . Then $(2z)^2 = 2y^2$, or $4z^2 = 2y^2$, or $2z^2 = y^2$, hence y^2 is even, and y should be even as well. But if x, y are both even, they have a common divisor 2, contradicting the assumption that they are relatively prime.

Analysis of Divisibility by Other Primes

Ok, what about the equation $x^2 - 17y^2 = 0$? It has real solutions (for example, $x = \sqrt{17}$, $y = 1$), and an even-odd analysis does not lead to contradiction: if both x, y are odd, then x^2 is odd and $17y^2$ is odd, so it may well be that $x^2 = 17y^2$. However, we can reach a contradiction by studying divisibility by 17 instead of 2. Indeed, let (x, y) be a relatively prime solution. Then $x^2 = 17y^2$, so x^2 is divisible by 17, thus $x = 17z$ for integer z , then $y^2 = 17z^2$, so y is divisible by 17 as well, a contradiction.

Our next example is $x^2 - 27y^2 = 0$. Here, we cannot say that “because x^2 is divisible by 27, then so is x ”, $x = 9$ is a counterexample ($x^2 = 81$ is divisible by 27, but $x = 9$ is not). This is because 27 is not prime. However, if $x^2 = 27y^2$ then surely x is divisible by 3, let $x = 3z$. Then $(3z)^2 = 27y^2$, $z^2 = 3y^2$, so z is divisible by 3 as well. With $z = 3t$, we get $(3t)^2 = 3y^2$, or $3t^2 = y^2$, hence y is divisible by 3 as well, a contradiction.

In fact, in our initial example $x^2 - 2y^2 = 0$, we can reach a contradiction by studying divisibility by 3 as well. Let (x, y) be a relatively prime solution. Every integer can be written as either $3k$, or $3k + 1$, or $3k + 2$ for integer k . If, for example, $x = 3k + 1$ and $y = 3m + 1$, then $(3k + 1)^2 = 2(3m + 1)^2$, or $9k^2 + 6k + 1 = 18m^2 + 12m + 2$, or $3(3k^2 + 2k - 6m^2 - 4m) = 2 - 1 = 1$, which is a contradiction because $3(3k^2 + 2k - 6m^2 - 4m)$ is divisible by 3 while 1 is not. All other cases ($x = 3k + 1$, $y = 3m + 2$, and so on) lead to a contradiction in a similar way.

Solutions in the p -Adic Field and the Hasse Principle

Using this method, we can reach a contradiction for any a, b except for the case when $a = kn^2, b = -km^2$ for some integers k, n, m . But then $x = m, y = n$ is a non-trivial solution. For example, the equation $4x^2 - 9y^2 = 0$ has a solution $x = 3, y = 2$, and many others, see Fig. 7.9a. Thus, we have three possible cases.

- It can be that Eq. (7.15) has a non-trivial integer solution, like $4x^2 - 9y^2 = 0$.
- Equation (7.15) may have no non-trivial *real* solutions, like $x^2 + y^2 = 0$. Then, of course, there are no non-trivial integer solutions as well.
- We can analyse the hypothetical relatively prime solution (x, y) for being even or odd, or, more generally, for divisibility by some other prime p , and reach a contradiction as in cases $x^2 - 2y^2 = 0$, $x^2 - 17y^2 = 0$, and $x^2 - 27y^2 = 0$ above.

In (c), if it is possible to obtain a contradiction using divisibility by a prime p , mathematicians say that the equation *has no solution in the p -adic field*. What is important is that, for equation (7.15), there are no other cases: if such an equation has no non-trivial integer solutions, we can *always* prove this by using either method (b) or method (c). In this case, we say that the equation satisfies the *Hasse principle*. The famous Hasse–Minkowski theorem states that this principle holds for any quadratic equation in s variables in the form $\sum_{i=1}^s \sum_{j=1}^s a_{ij}x_i x_j = 0$, where x_1, x_2, \dots, x_s are variables, and a_{ij} are coefficients.

Cubic Equations and Systems

In general, the Hasse principle does not hold for cubic equations. For example, the equation

$$3x^3 + 4y^3 + 5z^3 = 0$$

has no non-trivial integer solutions, but we need a different method to prove this. Indeed, this equation has a lot of non-trivial real solutions (for example, $x = -\sqrt[3]{3}, y = z = 1$), and “divisibility analysis” does not work as well.

Somewhat surprisingly, the problem becomes easier if we add more variables. For example, Roger Baker [32] proved that the cubic equation $a_1x_1^3 + \dots + a_7x_7^3 = 0$ in 7 variables *always* has a non-trivial integer solution (x_1, x_2, \dots, x_7) , no matter what the values of the integer parameters (a_1, \dots, a_7) are. With more variables, we can even guarantee the existence of solutions for *systems* of cubic equations. For example, Vaughan [388] proved in 1977 that the system

$$a_1x_1^3 + a_2x_2^3 + \dots + a_sx_s^3 = b_1x_1^3 + b_2x_2^3 + \dots + b_sx_s^3 = 0 \quad (7.16)$$

always has a non-trivial integer solution (x_1, x_2, \dots, x_s) if $s \geq 16$.

The Hasse Principle for Systems of Two Cubic Equations

Vaughan's result is the best possible, because in 1966 Davenport and Lewis [108] constructed an example of a system (7.16) with 15 variables which they proved has no integer solutions using the “divisibility by 7” method. Their system is

$$\begin{cases} f(x_1, \dots, x_5) + 7f(x_6, \dots, x_{10}) + 49f(x_{11}, \dots, x_{15}) = 0, \\ g(x_1, \dots, x_5) + 7g(x_6, \dots, x_{10}) + 49g(x_{11}, \dots, x_{15}) = 0, \end{cases}$$

where $f(x_1, \dots, x_5) = x_1^3 + 2x_2^3 + 6x_3^3 - 4x_4^3$, and $g(x_1, \dots, x_5) = x_2^3 + 2x_3^3 + 4x_4^3 + x_5^3$. The proof says that every x_i can be written as either $x_i = 7k_i$, or $x_i = 7k_i + 1$, or $x_i = 7k_i + 2, \dots$, or $x_i = 7k_i + 6$ for some integer k_i , and then in each case obtains a contradiction. In mathematical terminology, this means that the system has no solution in the 7-adic field.

The following theorem, proved in [79], states that the same method *always* works for system (7.16) whenever the number of variables is $s \geq 13$.

Theorem 7.9 Suppose that $s \geq 13$, and that $a_1, \dots, a_s, b_1, \dots, b_s$ are integer coefficients. Then the pair of Eqs. (7.16) has a non-trivial integer solution if and only if it has a non-trivial solution in the 7-adic field.

In other words, if system (7.16) has no non-trivial integer solution, we can *always* prove this using the “divisibility by 7” method! The authors also noted that their result is the best possible, in the sense that it does not hold for $s \leq 12$.

Reference

J. Br  dern and T. Wooley, The Hasse principle for pairs of diagonal cubic forms, *Annals of Mathematics* **166**-3, (2007), 865–895.

7.10 An Effective Multidimensional Szemer  di Theorem

Colouring Half the Integers Without Monochromatic Triples $(a, \frac{a+b}{2}, b)$

Assume that half of the positive integers are coloured red. Can you always find red integers a and b such that $\frac{a+b}{2}$ is also a red integer?

First, we need to clarify what exactly is meant by “half of the integers”? After all, there are infinitely many of them, so what is half of infinity? For the purpose of this discussion, let us assume that if we take the first N positive integers, then at least $N/2$ of them are red. For example, out of the first $N = 10$ integers, at least 5 are red. If these 5 integers are, for example, 1, 2, 5, 8, 10, then we can take red integers $a = 2$ and $b = 8$ such that $\frac{a+b}{2} = 5$ is also red. However, if the 5 red integers are 1, 2, 4, 5, 10, then, for any red a and b , $\frac{a+b}{2}$ is either not an integer, or not red.

Can we continue in this way and colour half of *all* integers without creating a red triple $(a, \frac{a+b}{2}, b)$? In fact, we cannot, and we would get stuck already for $N = 18$. You can colour 8 integers between 1 and 18 in red without creating the red triple above (for example, 1, 2, 4, 5, 10, 11, 13, 14) but not 9. For example, adding 15 to the above “red set” would result in the red triple 5, 10, 15, adding 16 would result in 4, 10, 16, adding 17 in 11, 14, 17, adding 18 in 10, 14, 18. Any other way to colour 9 numbers also does not work. Of course, if we cannot colour even the first $N = 18$ integers, there is no hope of colouring all of them in this way.

Colouring a Smaller Fraction of the Integers

It looks like half of the integers is too much, but what if we colour only one third of them? That is, 6 red out of the first 18, 7 red out of the first 21, and so on. It turns out that this works up to $N = 54$, for which we can colour $N/3 = 18$ integers 1, 2, 5, 6, 12, 14, 15, 17, 21, 31, 38, 39, 42, 43, 49, 51, 52, 54 with no red triple $(a, \frac{a+b}{2}, b)$. However, computer experiments show that, for $N = 57$, whatever $N/3 = 19$ integers are coloured red, we can always find a red triple $(a, \frac{a+b}{2}, b)$.

Maybe one third is also too much, what if we colour just 1% of all integers red? That is, just one out of the first 100, just 2 out of the first 200, and so on. In this case it seems to be very easy to avoid any red triple $(a, \frac{a+b}{2}, b)$. However, a famous theorem of Roth [326], proved in 1953, states that, for any $\delta > 0$, no matter how small, there exists an N such that if δN integers out of the first N are coloured red, we can always find a red triple in the form $(a, \frac{a+b}{2}, b)$.

Avoiding Longer Arithmetic Progressions

With $d = \frac{a-b}{2}$, the triple $(a, \frac{a+b}{2}, b)$ can be rewritten as $(a, a+d, a+2d)$, and it is called an arithmetic progression of length 3 with initial term a and difference d . More generally, a set $(a, a+d, a+2d, \dots, a+(k-1)d)$ is called an arithmetic progression of length k . Can we design reasonably large red sets which avoid at least longer progressions? This seems easier, for example, for $N = 38$, we can colour $N/2 = 19$ integers in red with no red 4-term arithmetic progression $(a, a+d, a+2d, a+3d)$. The 19 red numbers can be, for example,

$$1, 2, 3, 5, 6, 8, 9, 10, 15, 16, 17, 19, 26, 27, 29, 30, 31, 34, 37.$$

This sequence was obtained using a simple “greedy” algorithm, which takes numbers 1, 2, 3, ... and colours them red if there are no contradictions. Hence, 1, 2, 3 are red, but not 4 because this would create a red progression (1, 2, 3, 4). Then, 5 and 6 are red, but not 7 because of the progression 1, 3, 5, 7, and so on. After 37, the algorithm will not colour 38 because of the progression 26, 30, 34, 38, and not 39 because of 9, 19, 29, 39, and not 40 because of 31, 34, 37, 40... Hence, at least with this algorithm, only 19 numbers out of the first 40 can be coloured, which is less than half.

In fact, using smarter colouring algorithms, or colouring 1% of numbers instead of 50%, or avoiding even longer progressions, would only help for some initial values of N but not more. The celebrated Szemerédi theorem [365], proved in 1975, states that, for any natural number k , no matter how large, and for any $\delta > 0$, no matter how small, there exists an N such that if δN integers out of the first N are coloured red, we can always find a k -term arithmetic progression $(a, a + d, a + 2d, \dots, a + (k - 1)d)$, with all k terms red.

The Multidimensional Szemerédi Theorem

Geometrically, integers corresponds to certain points on the coordinate line, but every child can confirm that colouring a line is a bit boring, so let us do some colouring on the plane. For example, can we colour in red half of the points (n, m) with integer coordinates in such a way that we avoid a red square, that is, four red points with coordinates $(a, b), (a + d, b), (a + d, b + d), (a, b + d)$? Again, by “half of the points are red” we mean that, out of N^2 points in the “grid” $\{1, 2, \dots, N\}^2 := \{(n, m) \mid 1 \leq n \leq N, 1 \leq m \leq N\}$, at least $N^2/2$ should be red. In particular, for $N = 4$, we can easily find $N^2/2 = 8$ red points in $\{(n, m) \mid 1 \leq n \leq 4, 1 \leq m \leq 4\}$ which form no red square. For example, 12 points “on the boundary” of the square $((1, 1), (2, 1), (3, 1), (4, 1), (4, 2), (4, 3), (4, 4), (3, 4), (2, 4), (1, 4), (1, 3), (1, 2))$ forms only one square $((1, 1), (4, 1), (4, 4), (1, 4))$, hence, after excluding point $(4, 4)$, we find 11 points with the required property, see Fig. 7.10a. With the same method we can colour $15 > 5^2/2$ points for $N = 5$, and $19 > 6^2/2$ points for $N = 6$. However, for large N , only a tiny fraction of points are on the boundary. In fact, other colouring methods do not work as well, and, for large enough N , whichever $N^2/2$ points are coloured red, we can always find a red square. The same is true if we colour only δN^2 points for any fixed $\delta > 0$. Of course, there is nothing special about the square – the same statement remains true for triangles (for example, $\{(a, b), (a + 2d, b), (a, b + d)\}$), rectangles (say, $\{(a, b), (a + 3d, b), (a + 3d, b + d), (a, b + d)\}$), pentagons, and so on, see Fig. 7.10b).

In fact, the square $S = \{(a, b), (a + d, b), (a + d, b + d), (a, b + d)\}$ is just a unit square $X = \{(0, 0), (1, 0), (1, 1), (0, 1)\}$ enlarged d times and shifted by (a, b) , that is, $S = (a, b) + d \cdot X$. The two-dimensional Szemerédi theorem states that, for any set X of k points, and any $\delta > 0$, there exists an N such that in any colouring of δN^2 points, we can find a red set of the form $(a, b) + d \cdot X$. Moreover, the same is true if we colour integer points in three-dimensional space, and look for a red cube $S = \{(a, b, c), (a + d, b, c), (a + d, b + d, c), (a, b + d, c), (a, b, c + d), (a + d, b, c + d), (a + d, b + d, c + d), (a, b + d, c + d)\}$, or, more generally, for a set of the form $(a, b, c) + d \cdot X$ for any configuration X . Even more generally, one can colour points in \mathbb{Z}^r , that is, r -tuples (x_1, x_2, \dots, x_r) with all x_i integers, and prove the existence of various r -dimensional configurations which are all red.

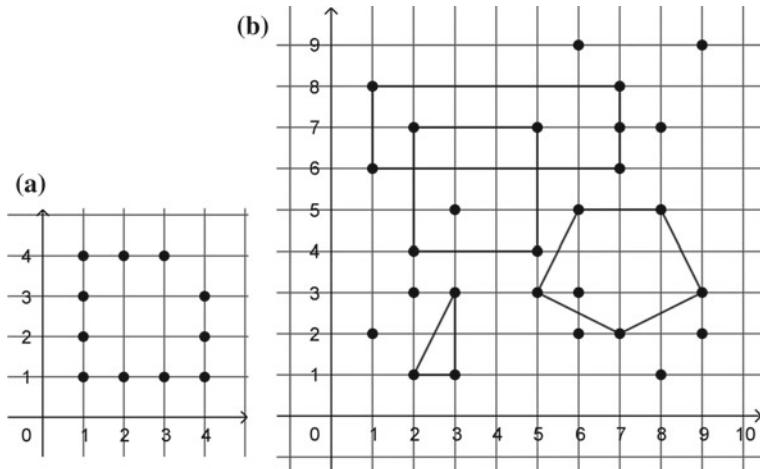


Fig. 7.10 Illustration of the two-dimensional Szemerédi theorem

Explicit Bounds for the Szemerédi Theorem

However, before 2007, all the known proofs of this result (called the multidimensional Szemerédi theorem [158]) provided no explicit bound for N . That is, we know that “for large enough N , whatever $N^2/2$ points in $\{1, 2, \dots, N\}^2$ are coloured red, we can always find a red square”, but cannot say exactly *how* large N should be for the theorem to hold. Would $N = 100$ suffice? Or a million? Or $N = 10^{10^{1,000,000}}$? There was no answer. Finally, in 2007, Timothy Gowers [168] gave the first proof which provides an explicit formula for N , as a function of δ , dimension r , and configuration X .

Theorem 7.10 *For every $\delta > 0$, every positive integer r , and every finite subset $X \subset \mathbb{Z}^r$, there is a positive integer N , explicitly computable from δ , r , and X , such that every subset A of the grid $\{1, 2, \dots, N\}^r$ of size at least δN^r has a subset of the form $a + dX$ for some $a \in \mathbb{Z}^r$ and integer $d > 0$.*

To prove Theorem 7.10, Gowers derived a generalization of one of the most important results in graph theory, Szemerédi’s regularity lemma. Recall that a graph $G = (V, E)$ is a set V of *vertices* and a set E of *edges*, which are 2-element subsets of V . Szemerédi’s regularity lemma states, roughly, that the vertices of every large graph can be divided into groups G_1, \dots, G_k of approximately equal size, such that the edges between different groups G_i and G_j are “approximately uniformly distributed”. That is, if we select any subset $A \subset G_i$ of size a , and $B \subset G_j$ of size b , then the density of edges between A and B (defined as the number of edges between A and B divided by ab) is approximately the same for almost every pair of A and B . To prove Theorem 7.10, Gowers derived a similar lemma for r -uniform hypergraphs, which are a generalization of graphs, where edges are r -element subsets of V (for

example, the set of vertices $V = \{1, 2, 3, 4\}$ with edges $(1, 2)$, $(2, 3)$, $(3, 4)$, and $(4, 1)$ is a graph, while $V = \{1, 2, 3, 4\}$ with edges $(1, 2, 3)$ and $(1, 2, 4)$ is a 3-uniform hypergraph). It is expected that this lemma should have more applications than Theorem 7.10 itself.

Reference

T. Gowers, Hypergraph regularity and the multidimensional Szemerédi theorem, *Annals of Mathematics* **166**-3, (2007), 897–946.

Chapter 8

Theorems of 2008



8.1 Minimal Surfaces in 3-Space II

From Congruency to Homeomorphism

In geometry, two objects or shapes are equal, or congruent, if one can be transformed into another using rigid motions, such as translations, rotations, or reflections. Intuitively, objects are equal if they have the same shape and size, for example, a square cannot be equal to a circle, two circles are equal if and only if they have the same radius, etc.

In contrast, in topology, two sets are considered to be equal, or *homeomorphic*, if they can be transformed into each other using continuous deformations, such as stretching and bending. For example, a square can be stretched to become rectangle, or bent to become a circle. More formally, subsets A and B of \mathbb{R}^3 are homeomorphic if there exists a one-to-one continuous function that transforms A into B and vice versa. For example, the function $f(x) = 2x$ continuously transforms the interval $(-1, 1)$ into the longer interval $(-2, 2)$. More generally, using the function $f(x) = \frac{b-a}{2}x + \frac{b+a}{2}$, the interval $(-1, 1)$ can be transformed into any interval (a, b) , hence all such intervals are homeomorphic. Moreover, the function $f(x) = \frac{x}{1-|x|}$ continuously transforms $(-1, 1)$ into the whole real line $(-\infty, \infty)$.

Transforming a Square into a Circle

Using “two-dimensional” continuous functions $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, like $f(x, y) = (2x, 2y)$, we can transform figures in the plane into each other. For example, $f(x, y) = (2x, 2y)$ transforms the square S with vertices $(1, 1), (1, -1), (-1, -1), (-1, 1)$ and side length 2 into the larger square with vertices $(2, 2), (2, -2), (-2, -2), (-2, 2)$ and side length 4. By a similar argument, all squares are homeomorphic to each other. More interestingly, the continuous function $f(x, y) = (2x, y)$ transforms the

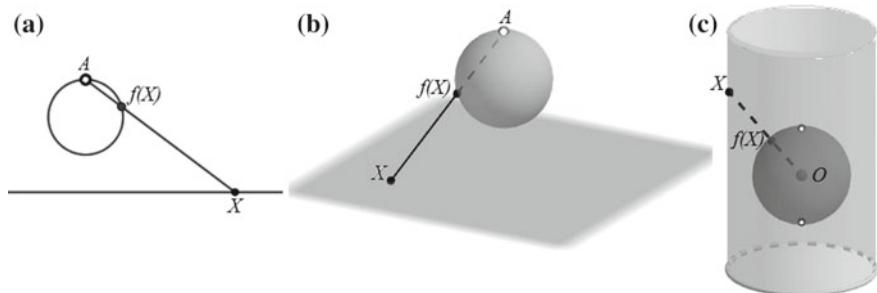


Fig. 8.1 Some examples of homeomorphisms

square S into the rectangle with vertices $(2, 1)$, $(2, -1)$, $(-2, -1)$, $(-2, 1)$, and in fact, all rectangles are homeomorphic. Moreover, let O be the coordinate center, C be any circle containing O , and X be any point on the square S . Then the ray OX intersects the circle C in a unique point which we call $f(X)$. The function f defined in this way is continuous, and transforms the square S into the circle C . In the same way a square can be transformed into a triangle, pentagon, ellipse, etc. So, all these objects are homeomorphic.

Transforming a Circle without a Point to a Line

However, a circle cannot be continuously transformed into an interval – these objects are fundamentally different even from a topological perspective. Intuitively, this is because in a circle (square, triangle) it is possible to go for a round trip, and return to the same point, while this is not possible inside the interval $(-1, 1)$. Interestingly, if we remove just one point from a circle, the round trip becomes impossible, and in fact, this “circle without a point” becomes homeomorphic to an interval. Indeed, let us draw the circle C on the coordinate plane with center $(0, 2)$ and radius 1. Let A be the point on C with coordinates $(0, 3)$, and $X = (x, 0)$ be any point on the x -axis. Then the line segment AX intersects the circle C at exactly one point, which we call $f(X)$, see Fig. 8.1a. Then f is a continuous function transforming the infinite line (x -axis) into the circle C without the point A . Hence, a “circle without a point” is homeomorphic to a line, which in turn is homeomorphic to $(-1, 1)$, as we have seen before.

Transforming Surfaces to a Disk

Let us now consider the area inside a circle, that is, a disk. An example is the unit disk B in the coordinate plane, which can be described as the set of points (x, y) satisfying the inequality $x^2 + y^2 < 1$. The function $f(x, y) = (2x, 2y)$ transforms it into a larger disk of radius 2. In the same way as we transformed a circle into a

square, we can transform the disk B into the interior of a square, triangle, etc. Next, in the same way as the function $f(x) = \frac{x}{1-|x|}$ transforms the interval $(-1, 1)$ into $(-\infty, \infty)$, the function $f(x, y) = \left(\frac{x}{1-\sqrt{x^2+y^2}}, \frac{y}{1-\sqrt{x^2+y^2}} \right)$ transforms the unit disk B into the whole infinite plane.

The disk B cannot be transformed into a sphere for reasons similar to why it is impossible to transform an interval into a circle. However, if we remove just one point from a sphere, this becomes possible, with essentially the same proof: take the sphere S with radius 1 and center $(0, 0, 2)$, select the point $A = (0, 0, 3)$ on it, and then for every point $X = (x, y, 0)$ on the x - y plane, let $f(X)$ be the point of intersection of the line segment AX with the sphere S , see Fig. 8.1b. Hence, a “sphere without a point” is homeomorphic to a plane, which in turn is homeomorphic to the unit disk B . In fact, topologists even omit the words “homeomorphic to”, and just say a “sphere without a point is a disk”, a “plane is a disk”, etc. Further, because $x^2 + y^2$ is a continuous function, the paraboloid defined by the equation $z = x^2 + y^2$ is also a disk, and so is any surface defined by an equation $z = f(x, y)$ for some continuous function f .

Transforming a Cylinder to a Sphere without Two Points

The surfaces of a cube, tetrahedron, etc., are homeomorphic to a sphere (in particular, they are not disks). The same is true for the surface of a finite cylinder. However, what about an infinite cylinder, that is, the set C of points (x, y, z) such that $x^2 + y^2 = 2$, z arbitrary. If S is the sphere with center $O = (0, 0, 0)$ and radius 1, and X is any point on the cylinder C , we can define $f(X)$ to be the point of intersection of the line segment OX with the sphere S , see Fig. 8.1c. Then f is continuous, but points $(0, 0, 1)$ and $(0, 0, -1)$ on the sphere do not correspond to any points of the cylinder. Hence, C is homeomorphic to the sphere S without *two* points. Another example of a surface which is not homeomorphic to a sphere is a torus.

Surfaces of Finite Topology

A set S is called *bounded* if the distance between any two of its points is less than some constant C . For example, the interval $I = (-1, 1)$, the ball $B = \{(x, y), x^2 + y^2 < 1\}$, and a sphere are bounded sets, while lines, planes, and infinite cylinders are not.

For the interval $I = (-1, 1)$, we can construct a sequence of points, for example, $1/2, 2/3, 3/4, \dots, n/(n+1), \dots$, which converges to a point outside the interval. The same is true for the unit disk $B = \{(x, y), x^2 + y^2 < 1\}$. On the other hand, if a sequence of points on a sphere converges to a limit, then this limit belongs to the sphere. Bounded surfaces with this property are called *compact*. The disk B can be

made compact if we add boundary points, that is, consider $B' = \{(x, y), x^2 + y^2 \leq 1\}$, but, interestingly, the sphere and torus are examples of compact surfaces without boundary points. Such surfaces are called *closed*. The disk B and cylinder are not closed surfaces, but are homeomorphic to a closed one (in this case, a sphere) without a finite number of points (in this case, without 1 or 2 points). Such surfaces are called *of finite topology*. This is an extremely broad class of surfaces: in fact, all the examples we considered here are of finite topology, and other examples exist but are difficult to visualize.

Minimal Surfaces and the Calabi Conjectures

An important class of surfaces, which frequently occur in nature, are *minimal* surfaces. Materials like soap bubbles form surfaces with minimal energy, which is equivalent to minimal area. Formally, a surface $M \subset \mathbb{R}^3$ is called a *minimal surface* if and only if every point $p \in M$ has a neighbourhood S with boundary B such that S has a minimal area out of all surfaces S' with the same boundary B . For example, a plane is a minimal surface, while a sphere is not, see Sect. 5.3 for more examples of minimal surfaces and a detailed discussion. Any subset of a minimal surface is again minimal, for example, a ball B is a minimal surface because it is a subset of the plane. A surface is called *complete* if it is not a proper subset of another (connected) surface. In 1965, Calabi [83] conjectured that any complete minimal surface in \mathbb{R}^3 must be unbounded, and, moreover, that any such surface, except for the plane, has an unbounded projection on every line. In general, both conjectures turn out to be false, but all known counterexamples have self-intersections (an example of a minimal surface with self-intersections is the Enneper surface, see Sect. 5.3). A surface without self-intersections is called *embedded*. The theorem below, proved in [97], confirms both Calabi conjectures for embedded minimal surfaces of finite topology.

Theorem 8.1 *The plane is the only complete embedded minimal surface with finite topology in a halfspace of \mathbb{R}^3 .*

In other words, if a complete embedded minimal surface S with finite topology is not a plane, then it cannot be covered by any halfspace. This of course implies that it is unbounded, and that it has an unbounded projection on every line.

Reference

T. Colding and W. Minicozzi, The Calabi–Yau conjectures for embedded surfaces, *Annals of Mathematics* **167**–1, (2008), 211–243.

8.2 Arbitrarily Long Arithmetic Progressions of Prime Numbers

Some Patterns in the Prime Numbers

The study of prime numbers, that is, numbers that have no divisors except for 1 and themselves, is almost as old as mathematics itself. The first primes are

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, \dots$$

We can see that there is a pair (2, 3) of consecutive primes. Is there at least one more such pair? No. In any pair of consecutive integers $(n, n + 1)$ one is odd and one is even. However, the only even prime is 2, hence the only pair of consecutive primes is (2, 3).

Ok, let us study further patterns. The primes (3, 5, 7) form a triple of primes at distance two from each other, that is, of the form $(n, n + 2, n + 4)$. Are there any more? After all, if n is odd, then $n + 2$ and $n + 4$ are odd as well, so the odd-even argument does not prevent them all from being primes. However, the problem is in divisibility by 3. Every integer n can be written as either $3k$, or $3k + 1$, or $3k + 2$. If $n = 3k$, then n is divisible by 3. If $n = 3k + 1$, then $n + 2 = 3k + 3 = 3(k + 1)$. If $n = 3k + 2$, then $n + 4 = 3(k + 2)$. In any case, either n , or $n + 2$, or $n + 4$ is divisible by 3. But 3 is the only prime divisible by 3, hence (3, 5, 7) is the only triple of primes in the form $(n, n + 2, n + 4)$.

The triple (3, 7, 11) is another interesting one, at distance 4 from each other, that is, of the form $(n, n + 4, n + 8)$. However, exactly the same proof shows that either n , or $n + 4$, or $n + 8$ is divisible by 3, hence (3, 7, 11) is the only triple of primes in this form.

3-Term Arithmetic Progressions in Primes

Further, the primes (5, 11, 17) are at distance 6 from each other. Is this triple unique again, similar to previous examples? It can be written as $(n, n + 6, n + 12)$, and if $n = 3k$, then n is divisible by 3 (and $n + 6$ and $n + 12$ are divisible by 3 as well). However, if $n = 3k + 1$, then $n + 6 = 3k + 7 = 3(k + 2) + 1$, and $n + 12 = 3(k + 4) + 1$ – so, none of these numbers are divisible by 3, and we have obtained no contradiction. Perhaps the problem is in divisibility by another number, say, 5? Any integer n can be written as either $n = 5k$, or $5k + 1$, or $5k + 2$, or $5k + 3$ or $5k + 4$. If $n = 5k$, then n can be prime only if $n = 5$, leading to our triple (5, 11, 17). If, for example, $n = 5k + 4$, then $n + 6 = 5k + 10 = 5(k + 2)$ is divisible by 5 and hence not prime. However, if, say, $n = 5k + 1$, then $n + 6 = 5k + 7 = 5(k + 1) + 2$, and $n + 12 = 5k + 13 = 5(k + 2) + 3$, so it may be that none of the numbers $(n, n + 6, n + 12)$ are divisible by 5. In a similar way we can prove that, for any prime p , it may be that none of the numbers $(n, n + 6, n + 12)$ are divisible by p . Hence, nothing prevents

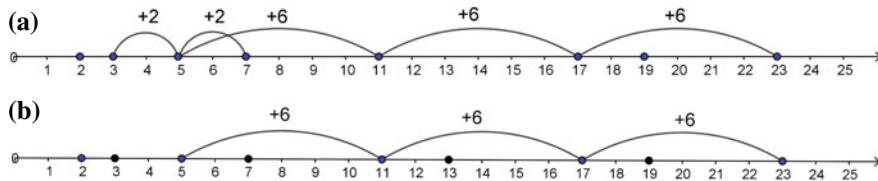


Fig. 8.2 Arithmetic progressions in **a** the primes, **b** a subset of the primes

them all from being primes. Indeed, we can find many such all-prime triples, for example, $(7, 13, 19)$, $(11, 17, 23)$, $(17, 23, 29)$, and so on.

In fact, such examples will never end. In general, a triple of integers at distance d from each other, that is, of the form $(n, n + d, n + 2d)$, is called an *arithmetic progression* of length 3. In 1939, van der Corput proved that there are infinitely many integers n and d such that n , $n + d$, and $n + 2d$ are all primes.

The Search for Larger Arithmetic Progressions

What if we look for longer patterns? For example, $(5, 11, 17, 23)$ is a sequence of four primes such that the difference between one term and the next is always 6 see Fig. 8.2a. Moreover, $(5, 11, 17, 23, 29)$ is a sequence of five primes with this property. In general, an arithmetic progression of length k is a sequence of the form $(n, n + d, n + 2d, \dots, n + (k - 1)d)$, where n and $d \neq 0$ are integers. What lengths of arithmetic progressions of primes we can find? By 2008, the record was $k = 23$. Frind et al. observed in 2004 that the 23 numbers

$$56211383760397 + 44546738095860 \cdot m, \quad m = 0, 1, \dots, 22,$$

are all primes.¹ Of course, no-one ever believed that this is the longest possible. In fact, there is an old folklore conjecture that we can find as long a progression as we like:

(C1) For every k , the prime numbers contain an arithmetic progression of length k .

Interestingly, conjecture (C1) immediately implies the following result, which sounds much stronger.

(C2) For every k , the prime numbers contain *infinitely many* arithmetic progressions of length k .

Why does (C1) imply (C2)? Well, by contradiction, assume that for some k , the prime numbers contain only finitely many (say, m) progressions of length k . However, by (C1), the primes contain at least one progression of length $K = k(m +$

¹At the time this book was published, the longest known *explicit* progression of primes has $k = 26$ terms.

1), let it be $n, n + d, n + 2d, \dots, n + (K - 1)d$. But then the primes contain $m + 1$ arithmetic progressions of length k , namely, $n, n + d, n + 2d, \dots, n + (k - 1)d$, and $n + kd, n + (k + 1)d, \dots, n + (2k - 1)d$, and so on, up to $n + kmd, n + (km + 1)d, \dots, n + (k(m + 1) - 1)d$. This is a contradiction.

A Heuristic Argument in Support of the Conjectures

Of course, (C2) also immediately implies (C1), so these conjectures are in fact equivalent. There is an extremely good reason to believe that they are true. The famous prime number theorem states that there are about $\frac{N}{\ln N}$ primes among the integers less than N . In other words, if we select an integer between 1 and N at random, then it will be a prime with probability about $\frac{N/\ln N}{N} = \frac{1}{\ln N}$. Now, let us fix some k and just select n and d at random, and check if $n, n + d, \dots, n + (k - 1)d$ all happen to be primes. Probability theory tells us that if the probability of some event is p , then the chance that it happens k times in independent experiments is p^k . In our case, $p \approx \frac{1}{\ln N}$, hence $p^k \approx \frac{1}{(\ln N)^k}$. For, say, $k = 23$, and large N (let us take $N = 10^{100}$), $\frac{1}{(\ln N)^k} \approx 5 \cdot 10^{-55}$. This chance is really tiny, which is why it is so hard to find long arithmetic progressions of primes in practice. However, in theory, we can assume that we have a *very* fast computer, which will try again and again, with new n and d . There are about N^2 possible choices for n and d , and, for $N = 10^{100}$, $N^2 = 10^{200}$. Of course, if you have 10^{200} trials, each will succeed with probability $5 \cdot 10^{-55}$, and it is difficult to imagine that you will not be “lucky” at least once. In fact, if you have M trials with probability of success p each time, you will expect to succeed about Mp times (for example, if you toss a coin $M = 1000$ times with a probability of $p = 1/2$ of getting Heads, you will expect to see about $Mp = 500$ Heads). Hence, one would expect that there exists not one, but about

$$C_k \frac{N^2}{(\ln N)^k}$$

arithmetic progressions of length k formed of prime numbers less than N , where C_k is a constant depending on k .

The Proof

However, this heuristic argument is not a proof, and, despite all efforts, conjecture (C2) was open for centuries even for $k = 4$. In 2008, Ben Green and Terence Tao [175] proved this conjecture in full for all k .

Theorem 8.2 *The prime numbers contain infinitely many arithmetic progressions of length k for all k .*

In fact, Green and Tao proved an even stronger result. Imagine that you colour every second prime blue, that is, the blue primes are 2, 5, 11, 17, 23, ..., see Fig. 8.2b.

Or every third prime. Or one out of every 100. Or one out of every L , you name your favourite large constant L . Moreover, you are allowed to choose which primes to colour, the only requirement being that, on average, every L -th prime should be coloured blue. Then, no matter how you do this, the *blue* prime numbers will always contain infinitely many arithmetic progressions of length k for all k !

Reference

B. Green and T. Tao, The primes contain arbitrarily long arithmetic progressions, *Annals of Mathematics* **167**-2, (2008), 481–547.

8.3 On Poincaré's Inequality

Convex Functions

A function f is called *convex* if

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) \quad (8.1)$$

for all x, y and for all $\alpha \in [0, 1]$. Geometrically, this means that if we select any points A and B on the graph of the function, the line segment AB lies above the graph, see Fig. 8.3a. Maybe the most well-known example of a convex function is $f(x) = x^2$. Indeed, in this case (8.1) reduces to $(\alpha x + (1 - \alpha)y)^2 \leq \alpha x^2 + (1 - \alpha)y^2$, which simplifies to $0 \leq \alpha(1 - \alpha)x^2 - 2\alpha(1 - \alpha)xy + (1 - \alpha)\alpha y^2 = \alpha(1 - \alpha)(x - y)^2$.

The definition (8.1) of a convex function can be equivalently written for n variables

$$f(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n) \leq \alpha_1 f(x_1) + \alpha_2 f(x_2) + \cdots + \alpha_n f(x_n), \quad (8.2)$$

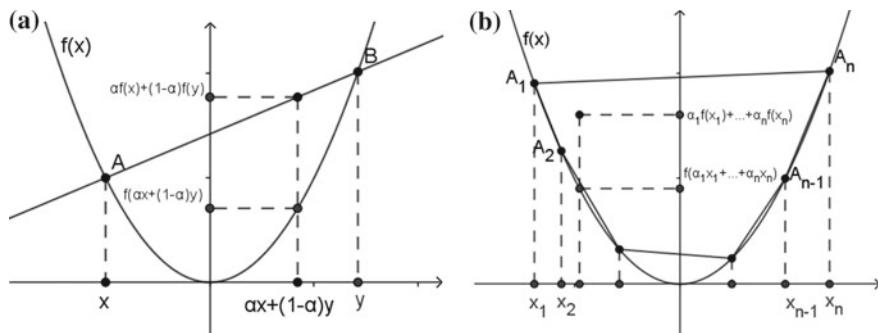


Fig. 8.3 The definitions of a convex function

whenever $\alpha_1 + \dots + \alpha_n = 1$. This can be easily proved by induction. Geometrically, it means that if we select any n points A_1, A_2, \dots, A_n on the graph of the function, the polygon $A_1 A_2 \dots A_n$ lies above the graph, see Fig. 8.3b.

Some Interesting Inequalities

With $\alpha_1 = \alpha_2 = \dots = \alpha_n = \frac{1}{n}$, (8.2) simplifies to

$$f\left(\frac{x_1 + x_2 + \dots + x_n}{n}\right) \leq \frac{f(x_1) + f(x_2) + \dots + f(x_n)}{n}, \quad (8.3)$$

that is, the function of the average value of x_i is less than or equal to the average value of $f(x_i)$. For the function $f(x) = x^2$, this implies that $\frac{x_1 + x_2 + \dots + x_n}{n} \leq \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}}$. More generally, for any $1 \leq p \leq q$ and non-negative y_i , one can prove that

$$\left(\frac{y_1^p + y_2^p + \dots + y_n^p}{n}\right)^{1/p} \leq \left(\frac{y_1^q + y_2^q + \dots + y_n^q}{n}\right)^{1/q}. \quad (8.4)$$

This follows from (8.3) with the function $f(x) = x^\beta$, $\beta = q/p \geq 1$, and $x_i = y_i^p$, $i = 1, 2, \dots, n$. It is left to verify that the function $f(x) = x^\beta$ is convex for any $x \geq 0$ and $\beta \geq 1$. An easy way to check this is to verify the inequality $f''(x) \geq 0$, where $f''(x)$ is the second derivative of f at x . For $f(x) = x^\beta$, the derivative is $f'(x) = \beta x^{\beta-1}$, and the second derivative is $f''(x) = \beta(\beta-1)x^{\beta-2} \geq 0$.

By applying the general version (8.2) of the definition of a convex function instead of (8.3), we could also prove the weighted version of (8.4)

$$\left(\frac{w_1 y_1^p + w_2 y_2^p + \dots + w_n y_n^p}{w_1 + w_2 + \dots + w_n}\right)^{1/p} \leq \left(\frac{w_1 y_1^q + w_2 y_2^q + \dots + w_n y_n^q}{w_1 + w_2 + \dots + w_n}\right)^{1/q}, \quad (8.5)$$

which holds for any non-negative weights w_1, w_2, \dots, w_n .

A Continuous Version of Inequality (8.5)

There is also a continuous version of inequality (8.5), written in terms of integrals. Intuitively, integration is a continuous version of summation. Geometrically, the integral $\int_I f(x) dx$ of a non-negative function x represents the area below the graph of $f(x)$ on an interval I . A continuous version of inequality (8.5) is

$$\left(\frac{\int_I y(x)^p w(x) dx}{\int_I w(x) dx}\right)^{1/p} \leq \left(\frac{\int_I y(x)^q w(x) dx}{\int_I w(x) dx}\right)^{1/q}, \quad 1 \leq p \leq q, \quad (8.6)$$

which is valid for non-negative functions $y(x)$ and $w(x)$, for which both integrals exist.

With $w(x) = 1$, the expression $\int_I w(x)dx$ simplifies to $\int_I dx$, which is just the length $|I|$ of the interval I (indeed, if $I = (a, b)$, then $\int_a^b dx = b - a = |I|$). In general, we will interpret $\int_I w(x)dx$ as a *weighted* length $|I|_w$ of the interval I . For example, if $w(x) = x^2$, then the weighted length of the interval $I = (0, 1)$ is $|I|_w = \int_0^1 x^2 dx = 1/3$.

The Weighted Average of a Function and Deviation from It

Further, with $w(x) = 1$, the expression $\frac{\int_I f(x)w(x)dx}{\int_I w(x)dx}$ simplifies to $\frac{1}{|I|} \int_I f(x)dx$, which is just the average value of f on the interval I . For example, the average value of the function $f(x) = x$ on $(0, 1)$ is $\frac{1}{1} \int_0^1 x dx = 1/2$. In general, we will interpret $\frac{\int_I f(x)w(x)dx}{\int_I w(x)dx}$ as a *weighted average* $f_{I,w}$ of f on I with respect to the weight w . For example, with weight $w(x) = x^2$, the average value of $f(x) = x$ on $I = (0, 1)$ is given by $f_{I,w} = \frac{1}{1/3} \int_0^1 x \cdot x^2 dx = \frac{3}{4}$. Then (8.6) just states that the p -th root of the weighted average of $y(x)^p$ increases if p increases.

In applications, it is important to estimate how much the function f deviates from its weighed average $f_{I,w}$. Define

$$\Delta(f, I, w) := \frac{1}{|I|_w} \int_I |f(x) - f_{I,w}|w(x)dx$$

to be the weighed average of the (absolute value of) the difference between f and $f_{I,w}$. In our example with $w(x) = x^2$, $f(x) = x$, and $I = (0, 1)$, we have $\Delta(f, I, w) = \frac{1}{1/3} \int |x - 3/4|x^2|dx = \frac{81}{512} \approx 0.16$.

Lipschitz Continuous Functions

Below, we learn how to estimate $\Delta(f, I, w)$ from above for so-called *Lipschitz continuous* functions. A function f is called Lipschitz continuous with constant L if $|f(x) - f(y)| \leq L \cdot d(x, y)$, $\forall x, y$, where $d(x, y) = |x - y|$ is the distance between x and y . For example, the functions $f(x) = |x|$, $f(x) = \sqrt{x^2 + 5}$, and $f(x) = \sin x$ are Lipschitz continuous with constant 1. In contrast, the function $f(x) = \sqrt{|x|}$ is not Lipschitz continuous on $[0, \infty)$, because, for $y = 0$ and $x > 0$, the inequality $|f(x) - f(y)| \leq L \cdot d(x, y)$ reduces to $\sqrt{x} \leq Lx$, or $x \leq (Lx)^2$, which fails for $x < 1/L^2$. For any Lipschitz continuous function f , define

$$\text{Lip } f(x) = \limsup_{y \rightarrow x} \frac{|f(x) - f(y)|}{d(x, y)}$$

(see Sect. 5.1 for the definition of \limsup). For example, for $f(x) = |x|$, $\text{Lip } f(x) = 1$ for all x , while if f is differentiable, then $\text{Lip } f(x) = |f'(x)|$ is just the absolute value of its derivative.

p -Admissible Weight Functions

All the above definitions can be easily generalized to functions $f(x_1, x_2, \dots, x_n)$ of many variables. The definition of convexity remains (8.1), but with $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$. Inequality (8.6) remains valid, but now $y : \mathbb{R}^n \rightarrow \mathbb{R}$ and $w : \mathbb{R}^n \rightarrow \mathbb{R}$ are functions of n variables, $I \subset \mathbb{R}^n$, and all integrals are n -dimensional. The definition of $\text{Lip } f(x)$ stays the same, but now $d(x, y)$ is a distance in \mathbb{R}^n . Further, for any $x \in \mathbb{R}^n$ and $r > 0$ let $B(x, r) = \{y \in \mathbb{R}^n \mid d(x, y) < r\}$ denote the ball with center x and radius r . A weight function $w : \mathbb{R}^n \rightarrow \mathbb{R}$ is called *p -admissible*, $p \geq 1$, if (i) there is a constant $C \geq 1$ such that $|B(x, 2r)|_w \leq C|B(x, r)|_w$ holds for all $x \in \mathbb{R}^n$ and $r > 0$, (ii) $0 < |B|_w < \infty$ for every ball B , and (iii) there are constants $C' \geq 1$ and $0 < t \leq 1$ such that the inequality

$$\Delta(f, B(x_0, tr), w) \leq 2rC' \left(\frac{1}{|B|_w} \int_B (\text{Lip } f(x))^p w(x) dx \right)^{1/p}$$

holds for all balls $B = B(x_0, r)$, and for every Lipschitz continuous function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ (this is called the $(1, p)$ -Poincaré inequality). p -admissible weights are important in the study of so-called “degenerate elliptic equations”.

Describing the Set of p Such That a Given Weight Function w is p -Admissible

Inequality (8.6) implies that if w is p -admissible for some $p \geq 1$, then it is also q -admissible for all $q \geq p$. The following theorem, proved in [216], states that, if $p > 1$, then w is also q -admissible for some q a bit *smaller* than p .

Theorem 8.3 *Let $p > 1$ and let w be a p -admissible weight in \mathbb{R}^n , $n \geq 1$. Then there exists an $\varepsilon > 0$ such that w is q -admissible for every $q > p - \varepsilon$.*

Theorem 8.3 implies that the set of $p \geq 1$ such that a given weight function w is p -admissible is either an interval $[1, \infty)$, or an open interval of the form (p_0, ∞) for some $p_0 \geq 1$.

Reference

S. Keith and X. Zhong, The Poincaré inequality is an open ended condition, *Annals of Mathematics* **167**-2, (2008), 481–547.

8.4 Representing Matrices in $SL_2(\mathbb{Z}/p\mathbb{Z})$ Using a Small Number of Generators

Composition and Inverse of Linear Functions

If we need to calculate the value of the expression $2x^3$ for, say, $x = 3$, we first calculate $x^3 = 27$, and then multiply by 2 to get 54. Similarly, to estimate 3^{x^2} for $x = 2$, we first calculate $x^2 = 4$, and then $3^4 = 81$. In general, a function $h(x)$ is called a *composition* of functions f and g if $h(x) = g(f(x))$. For example, the composition of functions $f(x) = x^3$ and $g(x) = 2x$ is $h(x) = g(f(x)) = g(x^3) = 2x^3$, the composition of $f(x) = x^2$ and $g(x) = 3^x$ is $h(x) = 3^{x^2}$, etc.

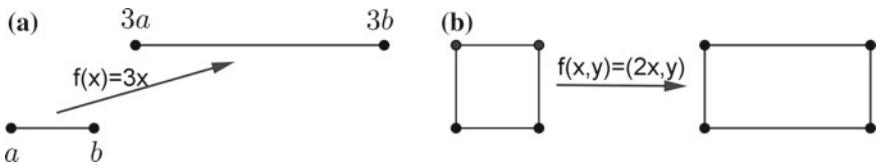
A function g is called the *inverse* function to f if $g(f(x)) = x$ for all x for which the functions are defined. In other words, if $f(x) = y$, then $g(y) = x$. For example, for $x \geq 0$, the inverse function to $f(x) = x^2$ is $g(x) = \sqrt{x}$, because $\sqrt{x^2} = x$.

Perhaps the simplest functions one can imagine are *linear* functions, that is, functions of the form $f(x) = ax + b$ for some coefficients a, b , like $f(x) = 2x$ or $f(x) = 3x + 1$. In this section, we consider only linear functions such that $f(0) = 0$. This implies that $f(0) = 0 \cdot a + b = 0$, hence $b = 0$, so that $f(x) = ax$. The composition of linear functions $f(x) = ax$ and $g(x) = bx$ is the function $h(x) = g(f(x)) = g(ax) = b \cdot ax = (ab)x$, which is also a linear function. Moreover, if the coefficients a and b are integers, then the coefficient ab of the function $h(x)$ is an integer as well. For example, the composition of $f(x) = 2x$ and $g(x) = 3x$ is $h(x) = g(2x) = 3 \cdot 2x = 6x$ – also a linear function with integer coefficient 6.

If $a \neq 0$, the inverse of the linear function $f(x) = ax$ is a linear function $g(x) = (1/a)x$. Indeed, $g(f(x)) = g(ax) = (1/a) \cdot (ax) = x$ for all x . However, unless $a = 1$ or $a = -1$, the inverse of a linear function with integer coefficient does *not* have an integer coefficient. For example, the inverse of the function $f(x) = 3x$ is $g(x) = (1/3)x$.

Linear Transformations of Lines and Planes

Geometrically, the function $f(x) = 3x$ describes a transformation of the coordinate line \mathbb{R} , for example, it “sends” point $x = 2$ to point $3x = 6$, interval $[-1, 1]$ into interval $[-3, 3]$, and, more generally, any interval $[a, b]$ into a larger interval $[3a, 3b]$, see Fig. 8.4a. Now, what if we need to describe a transformation of the coordinate *plane* which, for example, sends a square into a larger square, etc? For this, we need a function f which sends any point with coordinates (x, y) into another point with coordinates (u, v) . For example, the function $f(x, y) = (2x, 2y)$ transforms the square with vertex coordinates $(0, 0), (0, 1), (1, 1), (1, 0)$ into a larger square with side length 2, the function $f(x, y) = (2x, y)$ transforms this square into a rectangle with side lengths 1 and 2, see Fig. 8.4b, while the function $f(x, y) = (x, y)$ sends every point into itself, and is called the *identity function*.

**Fig. 8.4** Transformations of the coordinate line and plane

Such “two-dimensional” functions $f : (x, y) \rightarrow (u, v)$ can be arbitrary complicated, but, for simplicity, we consider only *linear* functions of the form $f(x, y) = (ax + by, cx + dy)$ for some coefficients a, b, c, d . Usually, mathematicians write these coefficients in the form of a 2×2 table $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, called a *matrix*, which we will denote by A_f . For example, the function $f(x, y) = (3x + 5y, 2x - 4y)$ corresponds to the matrix $A_f = \begin{bmatrix} 3 & 5 \\ 2 & -4 \end{bmatrix}$.

Function Composition and Matrix Products

For functions $f(x, y) = (2x, 2y)$, and $g(x, y) = (3x + 5y, 2x - 4y)$, their *composition* is

$$\begin{aligned} h(x, y) = g(f(x, y)) &= g(2x, 2y) = (3(2x) + 5(2y), 2(2x) - 4(2y)) \\ &= (6x + 10y, 4x - 8y). \end{aligned}$$

In general, the composition of functions $f(x, y) = (a_1x + b_1y, c_1x + d_1y)$, and $g(x, y) = (a_2x + b_2y, c_2x + d_2y)$ is

$$\begin{aligned} h(x, y) = g(f(x, y)) &= g(a_1x + b_1y, c_1x + d_1y) \\ &= (a_2(a_1x + b_1y) + b_2(c_1x + d_1y), c_2(a_1x + b_1y) + d_2(c_1x + d_1y)). \end{aligned}$$

The matrix A_h of coefficients of $h(x, y)$ is called the *product* of the matrices A_f and A_g of coefficients of f and g , and we write $A_f \cdot A_g = A_h$, or

$$\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} \cdot \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} = \begin{bmatrix} a_1a_2 + b_1c_2 & a_1b_2 + b_1d_2 \\ c_1a_2 + d_1c_2 & c_1b_2 + d_1d_2 \end{bmatrix}.$$

Inverse Matrices and Unimodularity

The inverse of a function $f(x, y) = (u, v)$ is a function g such that $g(u, v) = (x, y)$. For example, let us find the inverse of the function $f(x, y) = (2x + y, 3x + 2y)$. In

this case, $u = 2x + y$, $v = 3x + 2y$, and let us find x and y from this system. From the first equation, $y = u - 2x$. From the second one, $v = 3x + 2y = 3x + 2(u - 2x) = 2u - x$, hence $x = 2u - v$. Then $y = u - 2x = u - 2(2u - v) = -3u + 2v$. Hence, the function $g(u, v) = (2u - v, -3u + 2v) = (x, y)$ is the inverse to f . In general, the inverse to a function $f(x, y) = (ax + by, cx + dy)$ is given by $g(u, v) = \frac{1}{ad - bc}(du - bv, -cu + av)$, provided that $ad - bc \neq 0$. The corresponding matrix A_g is called the *inverse* of the matrix A_f and is usually denoted by A_f^{-1} . In general, the elements of A_f^{-1} are not integers even if all of a, b, c, d are integers, because of the division by $ad - bc$. Matrices with $ad - bc = 1$ are called *unimodular*, and the inverse of any unimodular matrix with integer elements also has integer elements. For the function $f(x, y) = (2x + y, 3x + 2y)$, the corresponding matrix $A_f = \begin{bmatrix} 2 & 1 \\ 3 & 2 \end{bmatrix}$ is unimodular (because $2 \cdot 2 - 1 \cdot 3 = 1$), that is why its inverse function g , computed above, has integer coefficients.

Arithmetic Modulo p

For the functions $f(x) = 13563x$ and $g(x) = 27695x$, can we quickly check if $h(x) = g(f(x))$ is divisible by 10 for $x = 382$? Yes, and the reason is that, to check divisibility by 10, we only need to keep the last digit in all calculations. $h(x) = 27695 \cdot 13563x$, and the last digit of $27695 \cdot 13563$ is 5 (because $5 \cdot 3 = 15$, and the last digit of 15 is 5), hence the last digit of $h(382)$ is 0 ($5 \cdot 2 = 10$, and the last digit of 10 is 0), and it is divisible by 10. The arithmetic in which we keep only the last digit, so that $3 \cdot 5 = 5$, $5 \cdot 2 = 0$, $6 + 5 = 1$, etc., is called *arithmetic modulo 10*. In fact, the last digit is just a remainder from division by 10, and, in a similar way, we could work with remainders from division by any other number. For example, in arithmetic modulo 7, we have $3 \cdot 5 = 1$ (because 15 gives remainder 1 when divided by 7), $5 \cdot 2 = 3$, $6 + 5 = 4$, etc. Similar rules apply to matrix arithmetic, for example, modulo 7,

$$\begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix} = \begin{bmatrix} 1 \cdot 1 + 3 \cdot 3 & 1 \cdot 0 + 3 \cdot 1 \\ 0 \cdot 1 + 1 \cdot 3 & 0 \cdot 0 + 1 \cdot 1 \end{bmatrix} = \begin{bmatrix} 3 & 3 \\ 3 & 1 \end{bmatrix}.$$

Note that all these matrices are unimodular modulo 7, the last one because $ad - bc = 3 \cdot 1 - 3 \cdot 3 = -6$, but $-6 = 1$ modulo 7. In general, the product of any two unimodular matrices is always unimodular. The set of all unimodular 2×2 matrices with multiplication defined modulo p for some prime p is denoted $SL_2(\mathbb{Z}/p\mathbb{Z})$.

Short Representations as a Product of Generators

While with the “usual” arithmetic there are infinitely many matrices, $SL_2(\mathbb{Z}/p\mathbb{Z})$ is a finite set. Indeed, there are only p possible remainders from division by p , that is,

only p possible values for matrix elements a, b, c, d . Hence, there are no more than p^4 matrices in $SL_2(\mathbb{Z}/p\mathbb{Z})$, and in fact less because not all of them are unimodular. Interestingly, for any p , any matrix $A \in SL_2(\mathbb{Z}/p\mathbb{Z})$ can be written as a product $A = G_1 \cdot G_2 \cdots \cdot G_k$, where each G_i is either $\begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix}$, or $\begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix}$, or the inverse of one of these two matrices. In general, a set $\mathcal{H} = \{H_1, \dots, H_m\}$ of matrices with this property (that is, such that every matrix A can be written as a product of matrices from \mathcal{H} and their inverses) is called a *set of generators*. A fundamental question is *how many* elements from \mathcal{H} we need for this representation. The following theorem of H. Helfgott [196] states that in fact we do not need many.

Theorem 8.4 *For every set of generators \mathcal{H} of $SL_2(\mathbb{Z}/p\mathbb{Z})$, every matrix $A \in SL_2(\mathbb{Z}/p\mathbb{Z})$ can be expressed as $A = G_1 \cdot G_2 \cdots \cdot G_k$, where $G_i \in \mathcal{H}$ or $G_i^{-1} \in \mathcal{H}$ for all i , and $k \leq M(\ln p)^c$, where M and c are some absolute constants.*

Reference

H. Helfgott, Growth and generation in $SL_2(\mathbb{Z}/p\mathbb{Z})$, *Annals of Mathematics* **167**-2, (2008), 601–623.

8.5 The Cayley Graphs of $SL_2(\mathbb{F}_p)$ Form a Family of Expanders

Graphical Representation of Multiplication Tables

If we need to determine whether $846732 \cdot 76177 + 90868$ is divisible by 10, there is no need to do the full calculation, it suffices to keep track of the last digit. The last digit of $846732 \cdot 76177$ is the same as the last digit of $2 \cdot 7$, which is 4. The last digit of 90868 is 8, and the last digit of $4 + 8$ is 2, hence $846732 \cdot 76177 + 90868$ is *not* divisible by 10 as it has remainder 2.

The “arithmetic” in which $2 \cdot 7 = 4$ and $4 + 8 = 2$ is called arithmetic modulo 10. Of course, modular arithmetic can be defined modulo any other integer. For example, the multiplication table for 2 by non-zero integers in arithmetic modulo 7 looks like

$$2 \cdot 1 = 2, \quad 2 \cdot 2 = 4, \quad 2 \cdot 3 = 6, \quad 2 \cdot 4 = 1, \quad 2 \cdot 5 = 3, \quad 2 \cdot 6 = 5.$$

This multiplication table can also be represented graphically. Let us represent the numbers 1, 2, 3, 4, 5, 6 as points in the plane, and connect any point a to its product $2 \cdot a$ by a line segment. That is, connect point 1 to point 2 (because $2 \cdot 1 = 2$), point 2 to point 4 (because $2 \cdot 2 = 4$), and point 4 to point 1 (because $4 \cdot 2 = 1$). For similar reasons, connect 3 to 6, 6 to 5, and 5 again to 3, see Fig. 8.5a. A set of points (called vertices) such that some of them are connected by a line (called an edge) is called a *graph*. In this case, our graph consist of two triangles, $1 - 2 - 4$ and $3 - 6 - 5$,

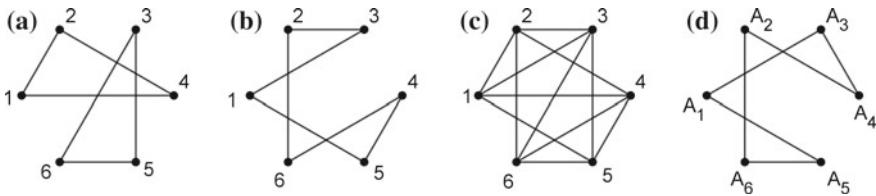


Fig. 8.5 Some examples of Cayley graphs

which are not connected to each other. From this graph, we immediately see that if we start from 1 and multiply by 2 arbitrarily many times, we will “move around the triangle $1 - 2 - 4$ ” and never reach, for example, 3. In “normal” arithmetic this means that 2^n never gives remainder 3 when divided by 7.

Can 3^n give remainder, say, 5, when divided by 7? To analyse this question, we can construct a similar graph based on multiplication by 3 modulo 7. That is, connect 1 to 3, 3 to 2 (because $3 \cdot 3 = 2$), 2 to 6, 6 to 4, 4 to 5, and 5 back to 1. In this case, the picture is not two disjoint triangles, but a connected hexagon with vertices 1, 3, 2, 6, 4, 5, see Fig. 8.5b. Starting from 1, we can visit any vertex. In particular, this means that 3^n can give any non-zero remainder when divided by 7, for example, 3^5 gives remainder 5.

How to Measure “How Connected” a Graph Is?

A graph is *connected* if any vertex is connected to any other one by a sequence of edges, or, equivalently, if the picture does not consist of two or more disjoint pieces. For example, a hexagon is a connected graph, while a graph consisting of two disjoint triangles is not. More generally, any n -gon is a connected graph.

In applications, it is important to understand “how much” the graph is connected. The connectivity can be “measured” by looking at how many edges need to be deleted to make the graph disconnected. For example, a hexagon can be made disconnected by deleting any two edges, and the same is true for any n -gon, so these graphs, while connected, are “not too connected”.

Formally, for any subset S of a graph, let $d(S)$ be the number of edges which connect S to the rest of the graph. For example, for our hexagon with vertices 1, 3, 2, 6, 4, 5 and $S = \{1, 3, 2\}$, S is connected to the rest of the graph by two edges (from 2 to 6 and from 1 to 5), so that $d(S) = 2$. Then, the *expansion* of S is $\frac{d(S)}{|S|}$, where $|S|$ is the number of vertices in S . In our example, $|S| = 3$, so that the expansion is $\frac{d(S)}{|S|} = \frac{2}{3}$. Finally, the *expansion coefficient* $c(G)$ of the whole graph G is the minimal possible expansion for all subsets S with size at most half of the size of the graph. In our example, if we take another S , for example, $S = \{1, 3\}$, then $d(S) = 2$ again, and $\frac{d(S)}{|S|} = \frac{2}{2} = 1 > \frac{2}{3}$. It is easy to check that no other S (with size at most half of size of the graph, that is, with $|S| \leq 3$) gives an expansion less than $\frac{2}{3}$, hence the expansion coefficient of the whole hexagon is $\frac{2}{3}$.

Families of Expanders

More generally, for an n -gon, any set S of consecutive vertices is connected to the rest of the n -gon by just two edges, so that $d(S) = 2$. If n is even and the size of S is $n/2$, this gives $\frac{d(S)}{|S|} = \frac{2}{n/2} = \frac{4}{n}$. If n becomes larger and larger, $\frac{4}{n}$ becomes closer and closer to 0. This means that a sequence of n -gons has low connectivity.

An infinite sequence of graphs $G_1, G_2, \dots, G_n, \dots$ is called a *family of expanders* if there is a constant $C > 0$ such that inequality

$$c(G_n) \geq C$$

holds for all graphs in the family, except for possibly finitely many exceptions. So, the sequence of n -gons is *not* a family of expanders. On the other hand, if G_n is a graph with n vertices such that every vertex is connected by an edge to every other vertex, then $c(G_n) = n/2$, so such a sequence $G_1, G_2, \dots, G_n, \dots$ is obviously a family of expanders. For applications, however, it is important to construct families of expanders with much fewer edges, see Sect. 2.3 for details and more examples.

“Joining” Graphs Representing Multiplication by 2 and 3

Returning to our modular arithmetic, we have constructed a graph with triangles $1 - 2 - 4$ and $3 - 6 - 5$ based on multiplication by 2 modulo 7, and a hexagon graph $1, 3, 2, 6, 4, 5$ based on multiplication by 3, see Fig. 8.5a, b. We can also “join” the above two graphs, that is, connect a to b if either $a = 2b$ or $a = 3b$ (or, vice versa, $2a = b$ or $3a = b$). In this case, 1 is connected to 2 and 4 (because $1 \cdot 2 = 2, 4 \cdot 2 = 1$) and also to 3 and 5 (because $1 \cdot 3 = 3, 5 \cdot 3 = 1$), and so on, see Fig. 8.5c. In general, for any prime p , and any set of “special” numbers $S = \{s_1, \dots, s_k\}$, we can construct a graph with $p - 1$ vertices, numbered from 1 to $p - 1$, and connect a to b if $a = s_i b$ (or $b = s_i a$) for some s_i .

Cayley Graphs of $SL_2(F_p)$

Similar graphs (called *Cayley graphs* with *set of generators* S) can be constructed from any finite set for which multiplication is well-defined (such sets are called *finite groups*). In particular, for any prime p , $SL_2(F_p)$ (or $SL_2(\mathbb{Z}/p\mathbb{Z})$) denotes the set of 2×2 tables of numbers (called *matrices*) $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, such that $ad - bc = 1$, with multiplication defined by

$$\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} \cdot \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} = \begin{bmatrix} a_1a_2 + b_1c_2 & a_1b_2 + b_1d_2 \\ c_1a_2 + d_1c_2 & c_1b_2 + d_1d_2 \end{bmatrix},$$

where all the operations are modulo p , see Sect. 8.4 for more details. For example, for $p = 2$, this is a 6-element set

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad A_5 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \quad A_6 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

If $S = \{A_3, A_5\}$, then the corresponding Cayley graph has 6 vertices, and the rules for edges are as above. For example, $A_1 \cdot A_3 = A_3$, $A_1 \cdot A_5 = A_5$, $A_2 \cdot A_3 = A_4$, $A_2 \cdot A_5 = A_6$, etc., hence A_1 is connected to A_3 and A_5 , A_2 is connected to A_4 and A_6 , and so on, see Fig. 8.5d. For any prime p , and any set of generators $S_p \subset SL_2(F_p)$, we can form such a Cayley graph G_p , and then $G_2, G_3, G_5, G_7, G_{11}, \dots$ is an infinite sequence of Cayley graphs. The following theorem of Jean Bourgain and Alex Gamburd [74] states that, if the sets S_p are selected at random, then this sequence is a family of expanders with probability 1.

Theorem 8.5 *Fix $k \geq 2$. For any p , let S_p be a k -element subset of $SL_2(F_p)$ whose elements are chosen independently at random, and let G_p be the corresponding Cayley graph. Then the sequence $G_2, G_3, G_5, G_7, G_{11}, \dots$ forms a family of expanders with probability 1.*

There are lots of mathematical results, theorems, and methods for studying families of expanders. With Theorem 8.5, all these results and methods can immediately be applied to study the properties of $SL_2(F_p)$.

Reference

J. Bourgain and A. Gamburd, Uniform expansion bounds for Cayley graphs of $SL_2(F_p)$, *Annals of Mathematics* **167**-2, (2008), 625–642.

8.6 Determining the Shape of an Infected Region

A (Very) Simple Model for Infection Spread

If, in a large city, several people are infected by a virus, how many people would become infected after time t , and how would the infection spread out geographically? Would infected people live more-or-less uniformly throughout the city, or will there be a region B such that almost all people inside B are infected, but almost all people outside B are healthy? If so, what is the expected size and shape of B ?

To answer these important but difficult questions, we need a mathematical model describing how infected people will move and infect other people. We start with a very simple model. Assume that (healthy) people live at points $0, 1, 2, 3, \dots$ of the coordinate line, and there is one infected person who is currently at point 0. He/she

will move from 0 to 1, then to 2, 3, and so on, infecting everyone he/she meets. After the first hour, he/she moves one step (from 0 to 1) with probability D , and stays at 0 with probability $1 - D$. The same process repeats after the second hour, and so on. The movements from n to $n + 1$ are called “jumps”. The movement is random, but we may use it to make some valuable predictions. For example, if $D = 1/4$, then how many people will be infected after $k = 100$ h? Intuitively, we would expect about $100 \cdot (1/4) = 25$ jumps, in which case 26 people would be infected (those who live at points 0, 1, ..., 25). In general, if the probability of a jump after each hour is D , then after k hours, we would expect about kD jumps (hence, about $kD + 1$ people infected).

This estimate holds only *on average*, and there is a (small) chance of significant fluctuations. For example, with probability $(1 - D)^k$, there will be no jumps at all during the first k hours, hence only one person (the one living at 0) will be affected.

A Continuous-Time Model and the Poisson Distribution

Of course, this simple model is very far from being realistic. First of all, it assumes that the jump may occur either after the first hour, or after the second one, and so on, while in reality time is continuous, and a person may decide to move at *any* moment, say, after 0.5 h, or after $\sqrt{2}$ h. To improve the model, we can assume that a person can move after every minute, but the probability of moving is $D/60$ (to keep the average jump rate at D jumps per hour). Or, even better, we can assume that the person can move after every second, with probability $D/3600$. More generally, we may allow jumps after each $1/N$ -th of an hour with probability D/N and let N go to infinity. In this case, the probability of no jumps after k hours becomes $(1 - D/N)^{Nk}$. With a new variable $N' = -N/D$, this expression becomes $(1 + 1/N')^{N' \cdot (-Dk)}$. However, $\lim_{N' \rightarrow \infty} (1 + 1/N')^{N'}$ is just a definition of the constant $e \approx 2.71828$, the base of the natural logarithm. Hence, in the limit, the probability of no jumps after t hours (we change notation from k to t to emphasize that t is now any real number, not an integer) should be

$$P(t, 0) = e^{-Dt}.$$

More generally, one can show that, in this model, the probability that after time t there will be exactly n jumps is given by

$$P(t, n) = e^{-Dt} \frac{(Dt)^n}{n!}, \quad n = 0, 1, 2, \dots$$

This is called the *Poisson distribution* with mean Dt .

Towards a More Realistic Model

Next, to make the model even more realistic, we can assume that (healthy) people may live at *all* integer points $-3, -2, -1, 0, 1, 2, 3, \dots$, and the affected person, who starts from 0, still moves with the average jump rate being D jumps per hour, but the jumps are to the right (from n to $n + 1$) or to the left (from n to $n - 1$) with equal chance. Even more generally, healthy people may live at all points (a, b) with integer coefficients, and the affected person, starting from $(0, 0)$, and making (on average) D jumps per hour, moves up, down, right, and left with equal chance. In other words, from point $x = (a, b)$ the jump may be one of points $(a + 1, b)$, $(a - 1, b)$, $(a, b + 1)$, or $(a, b - 1)$. This is called a *continuous time simple random walk on \mathbb{Z}^2 , with jump rate D* . In a similar way, we may define a random walk on \mathbb{Z}^d for any dimension d .

Next, we assume that, initially, there may be several (finitely many) infected people, and they may start their random walks from arbitrary positions, not necessarily from zero. Further, we assume that healthy people do not stay “at home”, but also move according to the same rules, independently of each other. And finally, let the initial number $N_0(x)$ of healthy people at any point $x \in \mathbb{Z}^d$ also be random, given by a Poisson distribution with some mean μ , and let $N_0(x)$ be independent from $N_0(y)$ if $x \neq y$. If any healthy person meets with any infected one, he or she becomes infected as well, and continues to move according to the same rules, infecting other people, and so on.

The Shape of the “Affected Region”

Now, let $A(t)$ be the set of all points visited by any infected person at any time between 0 and t . Because $A(t)$ is just a discrete set of points, it is difficult to talk about its “shape”. Let $B(t)$ be the set of points obtained by adding a unit cube around each point of $A(t)$. Formally, for $d = 2$, $(x, y) \in B(t)$ if and only if there is a point $(x^*, y^*) \in A(t)$ such that $|x - x^*| \leq 1/2$, and $|y - y^*| \leq 1/2$, and this definition can be easily extended to any dimension d . For example, if $A(t)$ is the set of 4 points $(0, 0)$, $(0, 1)$, $(1, 1)$, and $(1, 0)$, then $B(t)$ is just the square with vertices $(-1/2, -1/2)$, $(-1/2, 3/2)$, $(3/2, 3/2)$, $(3/2, -1/2)$, see Fig. 8.6a. In general, $B(t)$ serves as a reasonable model for the region in which people are affected by infection. Now, the question is how $B(t)$ grows with t .

Theorem 8.6 ([219]) *In the model described above, there exists a nonrandom, compact, convex set B_0 such that for all $\varepsilon > 0$, almost surely*

$$(1 - \varepsilon)B_0 \subset \frac{1}{t}B(t) \subset (1 + \varepsilon)B_0 \text{ for all large } t. \quad (8.7)$$

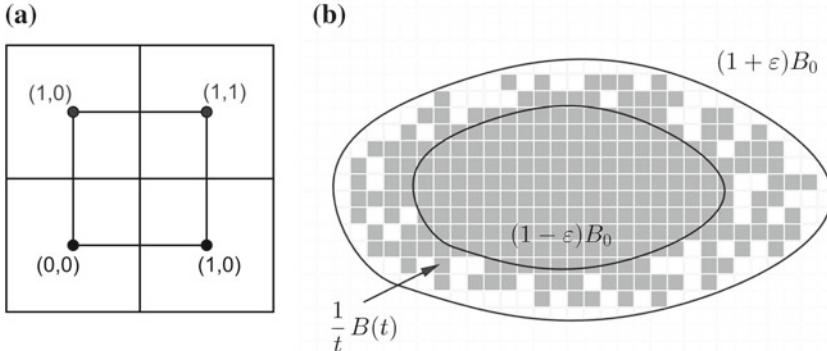


Fig. 8.6 The shape of the region affected by infection

See Fig. 8.6b for an illustration of the theorem. In case you are not familiar with the concepts involved in the formulation of Theorem 8.6, here are some relevant definitions and explanations. In d -dimensional space, a set B_0 is called *compact* if it bounded (that is, there is a point O and constant C such that the distance from O to any point $X \in B_0$ is at most C ; for example, circles and squares are bounded sets while a line is not) and closed (that is, contains its boundary points). A set B_0 is *convex* if it contains the line segment XY for any $X \in B_0$ and $Y \in B_0$.

In (8.7), $(1 - \varepsilon)B_0$ is, naturally, the set of all points x in the form $x = (1 - \varepsilon)y$ for some $y \in B_0$, and the sets $\frac{1}{t}B(t)$ and $(1 + \varepsilon)B_0$ are defined similarly. “For all large t ” in (8.7) means that, in almost every experiment, there exists a constant T (which depends on the experiment) such that the inclusions in (8.7) hold for all $t \geq T$. Finally, by saying “almost surely”, or “in almost every experiment” we mean “with probability 1”. In other words, (8.7) may fail, but the probability of this is equal to 0.

Intuitively, (8.7) means that, for large t , $B(t)$ looks like a scaled version of B_0 with “scale coefficient” t . This not only implies that the “affected region” $B(t)$ grows linearly with time, but also states that it converges to a particular shape. For example, it cannot be that the affected region looks like disk after a day, a square after two days, then again a disk, then a square, and so on. Hence, with Theorem 8.6, we have a reasonably precise understanding how the infection will spread out in time, at least if we believe in the model described.

Reference

H. Kesten and V. Sidoravicius, A shape theorem for the spread of an infection, *Annals of Mathematics* **167**-3, (2008), 701–766.

8.7 A Negative Answer to Littlewood's Question About Zeros of Cosine Polynomials

The Cosine Function

In a right triangle ABC with right angle A , the *cosine* of the angle B is the ratio of the length of the adjacent side to the length of the hypotenuse, that is $|AB|/|BC|$, see Fig. 8.7b. For example, it is easy to prove geometrically that in the triangle with angles $A = 90^\circ$, $B = 60^\circ$, $C = 30^\circ$ the length of AB is half of length of the hypotenuse BC . Hence, $\cos 60^\circ = 0.5$. In fact, it is more standard to measure angles in radians: 180° is π radians, hence 60° is $\pi/3$ radians. In this case, we write that $\cos(\pi/3) = 0.5$.

The above definition can be used to calculate the cosine only for acute angles, that is, angles between 0° and 90° , or, equivalently, between 0 and $\pi/2$ radians. However, there is a trigonometric formula $\cos(\pi/2 + x) = -\cos(\pi/2 - x)$ for all x , which can be used to calculate the cosine for obtuse angles, for example, $\cos(2\pi/3) = \cos(\pi/2 + \pi/6) = -\cos(\pi/2 - \pi/6) = -\cos(\pi/3) = -0.5$. Also, $\cos(\pi/2 + 0) = -\cos(\pi/2 - 0)$, hence $2\cos(\pi/2) = 0$, or $\cos(\pi/2) = 0$. Further, the formulas $\cos(-x) = \cos(x)$ and

$$\cos(x + 2\pi) = \cos(x), \quad \forall x \tag{8.8}$$

allow us to compute $\cos(x)$ for any real number x . For example, $\cos(-\pi/2) = \cos(\pi/2) = 0$, $\cos(7\pi/3) = \cos(2\pi + \pi/3) = \cos(\pi/3) = 0.5$, etc.

The Simplest Equations Involving the Cosine

Let us solve some equations involving the cosine, the simplest one is $\cos(x) = 0$. As we have already seen, $\cos(-\pi/2) = \cos(\pi/2) = 0$. Next, $\cos(3\pi/2) = \cos(2\pi - \pi/2) = \cos(-\pi/2) = 0$, $\cos(5\pi/2) = \cos(2\pi + \pi/2) = \cos(\pi/2) = 0$, etc. In general,

$$\cos(\pi/2 + \pi k) = 0$$

for all integers k . Hence, $x_k = \pi/2 + \pi k$, $k \in \mathbb{Z}$ (\mathbb{Z} denotes the set of all integers) forms an infinite family of solutions to the equation $\cos(x) = 0$, see Fig. 8.7a. One can prove that this equation has no other solutions.

In general, a function f is called *periodic* with period T if $f(x + T) = f(x)$, $\forall x$. If f is periodic, and the equation $f(x) = 0$ has a solution x_1 , then $x_1 + T$, $x_1 + 2T$, and, more generally, $x_1 + kT$ for any integer k are also solutions. From this infinite family of the form $x_1 + kT$, there is always one and only one solution in the interval $[0, T)$. Hence, to solve the equation $f(x) = 0$, it is sufficient to find all solutions x_1, x_2, \dots, x_m in $[0, T)$, and then all other solutions can be written as $x_i + kT$ for some i and $k \in \mathbb{Z}$. By (8.8), cosine is periodic with period 2π , hence we can restrict

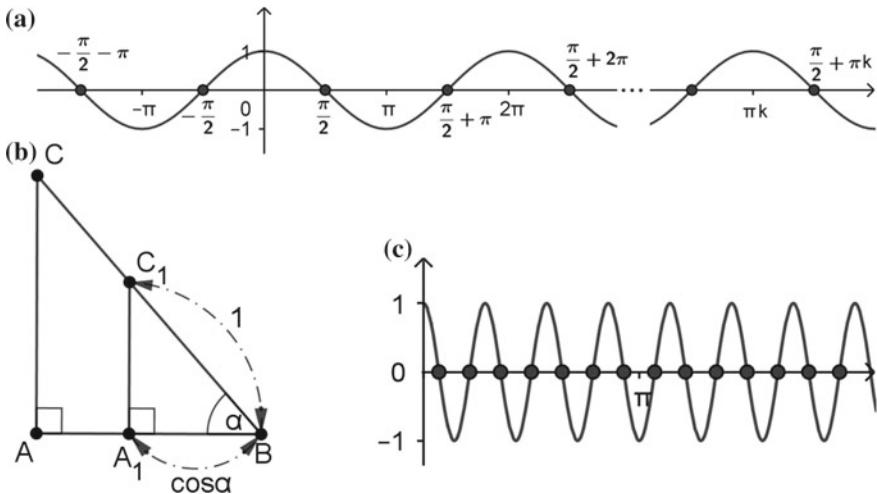


Fig. 8.7 a Cosine definition; b Graph and roots of $\cos(x)$; c Graph and roots of $\cos(nx)$

our attention to the interval $[0, 2\pi]$ only. On this interval, the equation $\cos(x) = 0$ has exactly two solutions, $x_1 = \pi/2$, and $x_2 = 3\pi/2$.

Let us solve the more general equation $\cos(nx) = 0$ for some integer n . In fact, because $\cos(-x) = \cos(x)$, we may assume that $n \geq 0$, since the equation with negative n is identical. Again, we are interested in solutions in the interval $[0, 2\pi]$ only. Because $\cos(nx) = 0$, $nx = \pi/2 + \pi k$ for some integer k . If $n \neq 0$, this implies that $x = \pi/2n + \pi k/n$. This solution belongs to the interval $[0, 2\pi]$ if $0 \leq \pi/2n + \pi k/n < 2\pi$, or $-1/2 \leq k < 2n - 1/2$. Because k is an integer, this implies that there are $2n$ solutions corresponding to $k = 0, 1, \dots, 2n - 1$, see Fig. 8.7c. If $n = 0$, $\cos(nx) = \cos(0) = 1 \neq 0$, hence the equation has no solutions.

An Equation with Two Cosines

Next, let us study the more general equation

$$\cos(n_1 x) + \cos(n_2 x) = 0,$$

where $n_1 \neq n_2$ are integers. There is a trigonometric formula stating that $\cos(\alpha) + \cos(\beta) = 2 \cos\left(\frac{\alpha+\beta}{2}\right) \cos\left(\frac{\alpha-\beta}{2}\right)$, which allows us to rewrite this equation as

$$2 \cos\left(\frac{n_1 + n_2}{2} x\right) \cos\left(\frac{n_1 - n_2}{2} x\right) = 0.$$

This is possible if either $\cos\left(\frac{n_1+n_2}{2}x\right) = 0$ or $\cos\left(\frac{n_1-n_2}{2}x\right) = 0$. In the first case, this implies that $\frac{n_1+n_2}{2}x = \pi/2 + \pi k$ for some integer k , which gives $|n_1 + n_2|$ solutions on $[0, 2\pi)$, while in the second case $\frac{n_1-n_2}{2}x = \pi/2 + \pi k$, which gives $|n_1 - n_2|$ solutions. Note that because $n_1 \neq n_2$, $|n_1 - n_2| \geq 1$, hence at least one solution is guaranteed to exist. In the case $n_1 = 0, n_2 = 1$, the equation $\cos(0) + \cos(x) = 1$ simplifies to $\cos(x) = -1$, and has *exactly* one solution in $[0, 2\pi)$, namely $x = \pi$.

An Equation with N Cosines, and a Question of Littlewood

A slightly more complicated argument shows that the equation

$$\cos(n_1x) + \cos(n_2x) + \cos(n_3x) = 0,$$

where n_1, n_2, n_3 are distinct non-negative integers, always has at least two solutions in $[0, 2\pi)$. In 1968, Littlewood [247] asked what is the minimal number of solutions in $[0, 2\pi)$ of the equation

$$\cos(n_1x) + \cos(n_2x) + \cdots + \cos(n_Nx) = 0, \quad (8.9)$$

where n_1, n_2, \dots, n_N are integers which are all different. For $N = 1, 2, 3$ the answer is 0, 1, 2, respectively, and Littlewood guessed that in general, the answer is “Possibly $N - 1$, or not much less”.

It turns out that this equation is indeed guaranteed to have at least $N - 1$ solutions for $N = 1, 2, \dots, 10$, but for $N = 11$ it is possible to find distinct non-negative integers n_1, n_2, \dots, n_{11} such that equation (8.9) has only 8 solutions. Also, for $N = 16$ the minimal number of solutions turns out to be 14. Hence, the “Possibly $N - 1$ ” part of Littlewood’s guess is wrong, but what about the “not much less” part?

Cosine Equations with Few Solutions

With the help of a computer, Borwein et al. found an example of an Eq. (8.9) with $N = 140$ which has only 52 solutions in $[0, 2\pi)$. Intuitively, 52 looks “much less” than $N - 1 = 139$, hence the “not much less” part of Littlewood’s guess seemed to be wrong as well. The following theorem, proved in [70], confirms that this is indeed the case for infinitely many values of N .

Theorem 8.7 *There exist a constant $C > 0$, a sequence of integers $N_1 < N_2 < \dots < N_m < \dots$, and cosine polynomials $\sum_{j=1}^{N_m} \cos(n_j x)$, where n_j are integers which are all different, such that the number of real solutions of (8.9) in $[0, 2\pi)$ does not exceed $C N_m^{5/6} \ln N_m$.*

For large N_m , $C N_m^{5/6} \ln N_m$ is much less than $N_m - 1$. In fact, the ratio $\frac{C N_m^{5/6} \ln N_m}{N_m - 1}$ converges to 0 as N goes to infinity. Hence, Eq. (8.9) can indeed have much fewer solutions than Littlewood expected.

Further Research

Even with Theorem 8.7, the original Littlewood question is far from being answered. If $S(N)$ is the minimal number of solutions to (8.9), then Theorem 8.7 just states that $S(N) \leq CN^{5/6} \ln N$ for infinitely many values of N , but this is just an upper bound, not an equality. To understand how big $S(N)$ really is, some lower bounds are required as well. However, even the proof that $S(N) \geq 1$ for $N \geq 2$ (that is, that (8.9) with $N \geq 2$ always has at least one solution) is not trivial. The fact that $S(N)$ goes to infinity as $N \rightarrow \infty$ was proved only recently by Tamás Erdélyi [128].

Reference

P. Borwein, T. Erdélyi, R. Ferguson, and R. Lockhart, On the zeros of cosine polynomials: solution to a problem of Littlewood, *Annals of Mathematics* **167**-3, (2008), 1109–1117.

8.8 The Kissing Number in Dimension Four is 24

The Kissing Number Question and Its Reformulations

Let S be a unit circle, that is, a circle of radius 1. How many non-overlapping unit circles S_1, S_2, \dots can you draw on the plane which all touch S ? The maximal number of such circles is called the *kissing number* in the plane.

Let O be the center of S . For any $i \neq j$, let O_i and O_j be the centres of circles S_i and S_j which touch S at points A_i and A_j , respectively, see Fig. 8.8a. Then S_i and S_j are non-overlapping if and only if the length $|O_i O_j|$ of the line segment $O_i O_j$ is at least 2. On the other hand, $|OA_i| = |A_i O_i| = |OA_j| = |A_j O_j| = 1$, because all circles have radius 1. Hence, in the triangle $OO_i O_j$, points A_i and A_j are the midpoints of sides AO_i and AO_j , respectively, which implies that $|A_i A_j| = \frac{1}{2}|O_i O_j|$, and the inequality $|O_i O_j| \geq 2$ is equivalent to $|A_i A_j| \geq 1$. Hence, our original question can be reformulated as: how many points A_1, A_2, \dots can you select on a unit circle S such that the distance between any two of them is at least 1?

Let A_i and A_j be any two points on the circle S . By the law of cosines for the triangle $OA_i A_j$,

$$|A_i A_j|^2 = |OA_i|^2 + |OA_j|^2 - 2|OA_i||OA_j| \cos(\angle A_i O A_j).$$

Because $|OA_i| = |OA_j| = 1$, this reduces to $|A_i A_j|^2 = 2 - 2 \cos(\angle A_i O A_j)$. Hence, $|A_i A_j| \geq 1$ if and only if $2 - 2 \cos(\angle A_i O A_j) \geq 1$, or $\cos(\angle A_i O A_j) \leq \frac{1}{2}$, which holds if the size of angle $\angle A_i O A_j$ is at least 60° , or $\pi/3$ radians. Hence, one more reformulation of the original question is: how many points A_1, A_2, \dots can you select on a circle S such that the size of the angle $\angle A_i O A_j$ is at least $\pi/3$, whenever $i \neq j$?

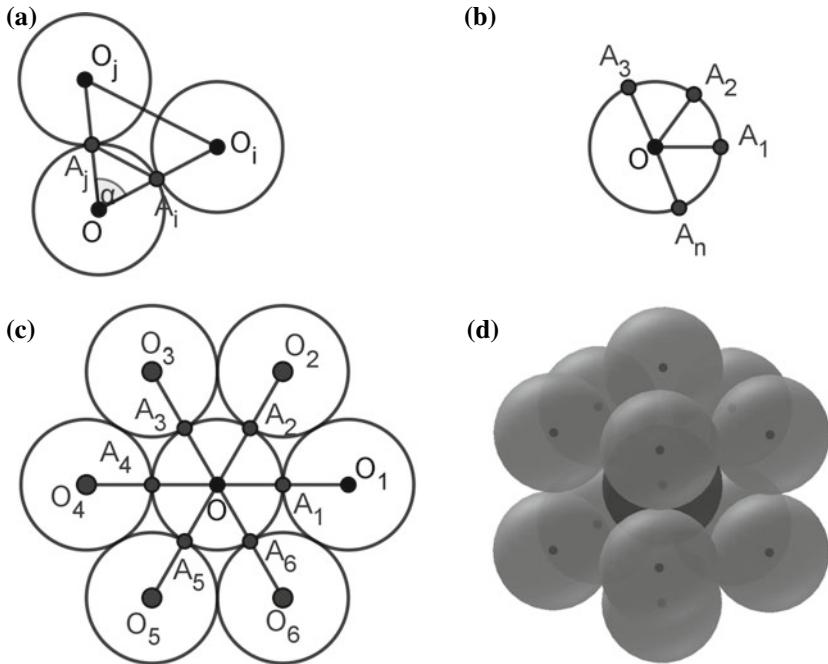


Fig. 8.8 The kissing number in dimensions 2 and 3

The Kissing Number in the Plane

In the last formulation, the question is easy. Assume that we have selected n points A_1, A_2, \dots, A_n , ordered clockwise. Then the angles $\angle A_1 O A_2, \angle A_2 O A_3, \dots, \angle A_n O A_1$ are all non-overlapping, and their sum is at most 2π , see Fig. 8.8b. Hence, if each of these angles is at least $\pi/3$, then $n \cdot \pi/3 \leq 2\pi$, which implies that $n \leq 6$. Equality holds if all angles $\angle A_1 O A_2, \angle A_2 O A_3, \dots, \angle A_6 O A_1$ have sizes exactly $\pi/3$, in which case the points A_1, A_2, \dots, A_6 form a regular hexagon with side length 1. The centres O_1, O_2, \dots, O_6 then form a regular hexagon with side length 2, see Fig. 8.8c. Hence, the kissing number in the plane is 6.

In fact, we can even generalize the question and ask how many points A_1, A_2, \dots, A_n we can select on a circle S such that $\angle A_i O A_j \geq \phi$, $i \neq j$, where $\phi \in (0, \pi)$ is an arbitrary angle, not necessarily $\pi/3$. This question is equally easy: a similar argument shows that the answer is the greatest integer n satisfying the inequality $n \leq \frac{2\pi}{\phi}$.

The Kissing Number in Three Dimensions

The problem becomes much more interesting in three dimensions. If S is a sphere of radius 1, how many non-overlapping spheres S_1, S_2, \dots of radii 1 exist in (three-dimensional) space which all touch S ? In other words, how many while billiard balls you can arrange around the black one (of the same size), which all touch it? The maximal number of spheres/billiard balls with this property is called the kissing number in three-dimensional space.

In the first reformulation, the problem becomes: how many points A_1, A_2, \dots exist on the sphere S with pairwise distance at least 1 between each other? In other words, how many cities you can build on Earth (assuming that Earth is sphere and cities are points) such that the distance between any two of them is at least the radius of the Earth? Equivalently, the angle $\angle A_i O A_j$ (where O is the center of the Earth/sphere) should be at least $\pi/3$ whenever $i \neq j$. Of course, one can generalize the problem by replacing $\pi/3$ with any angle ϕ , or, equivalently, ask how many cities we can build on Earth with pairwise distance at least d , where $d > 0$ is any fixed distance, say, 50 km.

It is easy to prove that the kissing number in three-dimensional space is at least 12. There is a nice symmetric construction for $n = 12$, where A_1, A_2, \dots, A_{12} are the vertices of a regular icosahedron, see Fig. 8.8d. In this case, we have $|A_i A_j| \geq 1/\sin(2\pi/5) \approx 1.05 > 1$, hence there is even some room to move points A_i while keeping all pairwise distances at least 1. This opens the possibility of moving all neighbours of A_1 closer to it, then all “neighbours of neighbours” closer to them, and so on, and finally trying to find the space for one more point, A_{13} , at distance at least 1 from each of A_1, A_2, \dots, A_{12} . This was the subject of a famous discussion between Isaac Newton and David Gregory in 1694, in which Gregory believed that this might be possible, while Newton believed that it is not. It turned out that Newton was right, but the proof was found only after more than 250 years, in 1952, by Schütte and van der Waerden [339].

Moving to Higher-Dimensional Spaces

Ok, so the kissing numbers in dimensions 2 and 3 are 6 and 12, respectively. Is this the end of the story? No, because the same question can be asked in higher dimensions. In dimension 4, every point X can be described using four real coordinates, (x_1, x_2, x_3, x_4) , with the distance between points $X = (x_1, x_2, x_3, x_4)$ and $Y = (y_1, y_2, y_3, y_4)$ defined as

$$\rho(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + (x_4 - y_4)^2}.$$

The unit sphere S with center O is the set of all points X such that $\rho(O, X) = 1$, and the kissing number is the maximal number of points A_1, A_2, \dots, A_n on S such that $\rho(A_i, A_j) \geq 1, i \neq j$. The same question can be asked for any dimension d .

Although we live in 3-dimensional space, research in higher dimensions is not just a purely theoretical exercise. In fact, a point X with coordinates (x_1, x_2, \dots, x_d) in d -dimensional space may represent any object X (for example, a car) with d parameters (for example, x_1 is the maximal speed, x_2 is the price, etc.). Then the kissing number represents the number of objects (cars) which are all “sufficiently different” from each other. This has some important applications, for example, in coding theory.

The Solution in Dimension Four

In dimension 4, there is a simple construction which proves that the kissing number is at least 24: take points with coordinates in the form $(\pm \frac{1}{\sqrt{2}}, \pm \frac{1}{\sqrt{2}}, 0, 0)$, where you can choose the signs and also permute the coordinates. There are 4 possible sign patterns $(++, +-, -+, --)$, and, for each of them, we have six permutations (for the $++$ pattern, they are $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0)$, $(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, 0)$, $(\frac{1}{\sqrt{2}}, 0, 0, \frac{1}{\sqrt{2}})$, $(0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)$, $(0, \frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}})$, and $(0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$), resulting in 24 points in total. These points all lie on the unit sphere with center $(0, 0, 0, 0)$, and it is easy to check that the distance between any two of them is at least 1.

In 1963, Coxeter [104] proposed a general method for proving upper bounds for kissing numbers in all dimensions. In dimension four, this method gives the upper bound 26. Later, Delsarte proposed a different method which, as demonstrated in [291], gives the upper bound 25. However, the question whether the construction with 25 points is possible remained open, until it was answered negatively in 2008 by the following theorem of Musin [287].

Theorem 8.8 *The kissing number in dimension four is 24.*

In other words, the above construction with 24 points is optimal in dimension 4.

What about even higher dimensions? Interestingly, the kissing number is known exactly in dimensions... 8 and 24. The reason is that, in these dimensions, there are particularly nice constructions of 240 and 196560 points, respectively, and it is possible to use Delsarte’s method to prove that this is optimal [291]. However, the problem remains open in all other dimensions. For example, in dimension 5, the best known construction contains 40 points, while the best upper bound is 44.

Reference

O. Musin, The kissing number in four dimensions, *Annals of Mathematics* **168**-1, (2008), 1–32.

8.9 A Criterion for Embedding L_p into L_q Uniformly

Several Ways to Measure the Distance Between Points

A standard way to measure the distance between points A and B on the plane is

$$d(A, B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2},$$

where (x_A, y_A) and (x_B, y_B) are the coordinates of A and B , respectively. For example, if $A = (0, 0)$ and $B = (1, 2)$, then $d(A, B) = \sqrt{(0-1)^2 + (0-2)^2} = \sqrt{5}$.

However, if A and B represent, say, buildings in a city, it is usually impossible to travel from A to B using a straight-line path, you need to use streets. For example, if the streets are parallel to the coordinate axes, a shortest path from $(0, 0)$ to $(1, 2)$ is $(0, 0) \rightarrow (1, 0) \rightarrow (1, 2)$, and its length is 3, see Fig. 8.9a. More generally,

$$d'(A, B) = |x_A - x_B| + |y_A - y_B|.$$

The distances $d(A, B)$ and $d'(A, B)$ correspond to the special cases ($p = 2$ and $p = 1$) of the more general formula

$$d_p(A, B) = \sqrt[p]{|x_A - x_B|^p + |y_A - y_B|^p}, \quad p \geq 1.$$

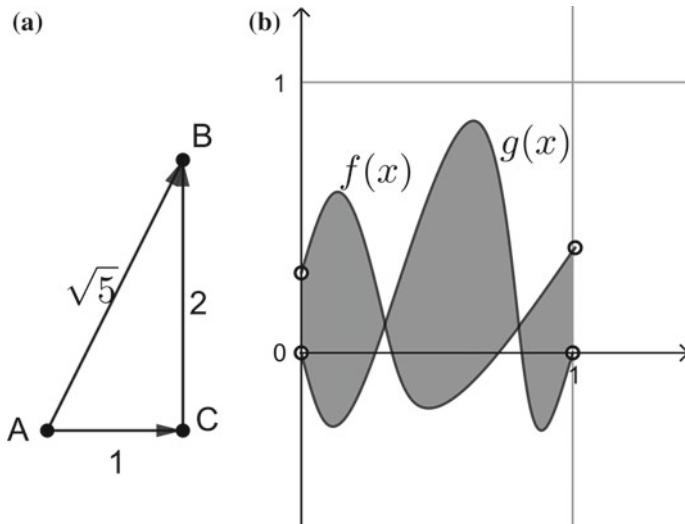


Fig. 8.9 Distance between points and functions

The Comparison Between Different Distances

Of course, $d(A, B) \leq d'(A, B)$: it would be shorter to travel from A to B using a straight-line path than using parallel-axis streets. However, it is not *much* shorter. If $A = (0, 0)$, and $B = (1, 1)$, then $d(A, B) = \sqrt{2}$ and $d'(A, B) = 2$, hence $\frac{d(A, B)}{d'(A, B)} = \frac{1}{\sqrt{2}} \approx 0.7$, but you can never have $d(A, B) < \frac{1}{\sqrt{2}}d'(A, B)$. Indeed,

$$\begin{aligned} (d'(A, B))^2 &= (x_A - x_B)^2 + (y_A - y_B)^2 + 2(x_A - x_B)(y_A - y_B) \\ &\leq 2(x_A - x_B)^2 + 2(y_A - y_B)^2 = 2d^2(A, B), \end{aligned}$$

where the inequality follows from $[(x_A - x_B) - (y_A - y_B)]^2 \geq 0$. Hence, $\frac{1}{\sqrt{2}}d'(A, B) \leq d(A, B) \leq d'(A, B)$. In particular, this implies that if $d(A, B)$ is “small” then so is $d'(A, B)$. More formally, for any $\varepsilon > 0$, however small, there exists a $\delta > 0$ such that $d(A, B) < \delta$ implies that $d'(A, B) < \varepsilon$. In fact, we can take $\delta = \frac{1}{\sqrt{2}}\varepsilon$.

Measuring the Distance Between Functions

More generally, one may wish to measure the distance not only between points, but also between arbitrary objects, for example, functions. For example, let f and g be any functions from $[0, 1]$ to \mathbb{R} for which the integrals $\int_0^1 |f(x)|dx$ and $\int_0^1 |g(x)|dx$ are well-defined and finite. Then define the distance

$$d_{(1)}(f, g) = \int_0^1 |f(x) - g(x)|dx.$$

Geometrically, $d_{(1)}(f, g)$ is the area of the shaded region in Fig. 8.9b. For example, the distance between $f(x) = x$ and $g(x) = x^2$ is $d_{(1)}(f, g) = \int_0^1 (x - x^2)dx = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$.

Of course, $d_{(1)}(f, g)$ is not the only way to measure the distance between function. By analogy with the “usual” distance $d(A, B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$ between points, we may define, for functions,

$$d_{(2)}(f, g) = \sqrt{\int_0^1 (f(x) - g(x))^2 dx},$$

whenever the integral exists and is finite, or, more generally,

$$d_{(p)}(f, g) = \sqrt[p]{\int_0^1 |f(x) - g(x)|^p dx}, \quad p \geq 1. \quad (8.10)$$

The Correspondence Between Points and Functions

In fact, we may associate to any point A of the plane a function f_A such that, informally speaking, “close” points A and B correspond to “close” functions f_A and f_B , and vice versa. Specifically, to point A with coordinates (x_A, y_A) , let us associate a function f_A such that $f_A(x) = x_A$ for $x \in [0, 1/2]$ and $f_A(x) = y_A$ for $x \in (1/2, 1]$. If the function f_B corresponds to the point B with coordinates (x_B, y_B) , then

$$d_{(1)}(f_A, f_B) = \int_0^1 |f_A(x) - f_B(x)| dx = \int_0^{1/2} |x_A - x_B| dx + \int_{1/2}^1 |y_A - y_B| dx = \frac{1}{2} d'(A, B).$$

In particular

- (a) for any $\varepsilon > 0$, there exists a $\delta > 0$ such that $d_{(1)}(f_A, f_B) < \delta$ implies that $d'(A, B) < \varepsilon$,

(in fact we can take $\delta = \varepsilon/2$), and, vice versa,

- (b) for any $\varepsilon > 0$ there exists a $\delta > 0$ such that $d'(A, B) < \delta$ implies that $d_{(1)}(f_A, f_B) < \varepsilon$,

(this time we can take $\delta = 2\varepsilon$).

Metric Spaces and Uniform Embeddings

Distance may be studied not only between points or functions, but between arbitrary abstract objects. In general, a *metric space* is any set S , and a function d which assigns to every pair $a, b \in S$ a real number $d(a, b)$, called a distance, such that (i) $d(a, b) \geq 0$, (ii) $d(a, b) = 0$ if and only if $a = b$, (iii) $d(a, b) = d(b, a)$, and (iv) $d(a, b) + d(b, c) \geq d(a, c)$, $\forall a, b, c$ (triangle inequality), see Sect. 5.8 for more details. For example, for any $p \geq 1$, the set of all points in the plane with distance $d_p(A, B)$ is a metric space. Also, the set of all functions $f : [0, 1] \rightarrow \mathbb{R}$ such that $\int_0^1 |f(x)|^p dx$ exists and is finite, and with distance defined as in (8.10) is a metric space, called L_p space, $p \geq 1$. For $p < 1$, $d_{(p)}(f, g)$ in (8.10) does not satisfy the triangle inequality, but another distance

$$d'_{(p)}(f, g) = \int_0^1 |f(x) - g(x)|^p dx, \quad p \in (0, 1)$$

satisfies (i)–(iv) and defines the metric space L_p for $p \in (0, 1)$.

A metric space X (with distance d_X) is said to be *uniformly embedded* into a metric space Y (with distance d_Y) if we can associate to every element $a \in X$ an element $h(a) \in Y$ such that $h(a) \neq h(b)$ whenever $a \neq b$, and

- (a) for any $\varepsilon > 0$, there exists a $\delta > 0$ such that $d_Y(h(a), h(b)) < \delta$ implies that $d_X(a, b) < \varepsilon$, and

- (b) conversely, for any $\varepsilon > 0$ there exists a $\delta > 0$ such that $d_X(a, b) < \delta$ implies that $d_Y(h(a), h(b)) < \varepsilon$.

In other words, a and b are “close” in X if and only $h(a)$ and $h(b)$ are “close” in Y . For example, as we saw above, the metric space of all points in the plane with distance $d'(A, B)$ can be uniformly embedded into the metric space of functions with distance $d_{(1)}(f_A, f_B)$, that is, into L_1 space.

Embedding L_p into L_q

Much less trivial is to uniformly embed one space of functions into another one. In particular, for which p and q can we embed L_p into L_q ? This question was open for decades, until it was answered by the following theorem of Manor Mendel and Assaf Naor [271].

Theorem 8.9 *For $p, q > 0$, L_p embeds uniformly into L_q if and only if $p \leq q$ or $q \leq p \leq 2$.*

In fact, Theorem 8.9 is just one of many corollaries of a nice and general theory. Mendel and Naor found a way to associate to every metric space S a positive real number $r(S)$, called the *metric cotype* of S , which has a lot of useful properties, from which Theorem 8.9, as well as many more nice results, follow immediately.

Reference

M. Mendel and A. Naor, Metric cotype, *Annals of Mathematics* **168**-1, (2008), 247–298.

8.10 The Distribution of Integers with a Divisor in a Given Interval

Integers Divisible by Either 3 or 4

Let x be an integer divisible by 12, for example, $x = 24$. How many integers less than or equal to x are divisible by either 3 or 4?

Well, there are $x/3$ integers divisible by 3, and $x/4$ integers divisible by 4, totalling in $x/3 + x/4 = 7x/12$. For $x = 24$, this gives $7 \cdot 24/12 = 14$. However, direct counting gives just 12 such integers: 3, 4, 6, 8, 9, 12, 15, 16, 18, 20, 21, 24. The problem is that, while there are indeed $24/3 = 8$ integers divisible by 3 (3, 6, 9, 12, 15, 18, 21, 24), and $24/4 = 6$ integers divisible by 4 (4, 8, 12, 16, 20, 24), there are two integers (12 and 24) which are divisible by 3 and 4 *simultaneously*, and therefore belongs to *both* these sets, see Fig. 8.10a. Hence, when we add 8 and 6 to get 14, we have counted these integers twice. To get the correct answer, we now need to subtract the number of integers counted twice, which results

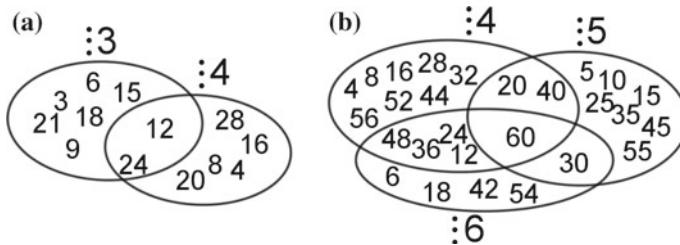


Fig. 8.10 Integers divisible by **a** 3 and 4, **b** 4, 5, and 6

in $14 - 2 = 12$. For general x (divisible by 12), the number of integers less than or equal to x which are divisible by both 3 and 4 is $x/12$. Hence, from $7x/12$, we need to subtract $x/12$ numbers counted twice, to get the correct answer: $x/2$.

Integers Divisible by Either 4, 5, or 6

Now, for x divisible by 60, how many integers less than or equal to x are divisible by either 4, 5, or 6? As above, there are $x/4$ integers divisible by 4, $x/5$ integers divisible by 5, and $x/6$ integers divisible by 6, totalling $x/4 + x/5 + x/6$. Again, integers divisible by 4 and 5 are counted twice and need to be subtracted. An integer is divisible by 4 and 5 if and only if it is divisible by 20, hence there are $x/20$ such integers. Similarly, integers divisible by 4 and 6 are counted twice. These are integers divisible by 12, hence there are $x/12$ of them. Finally, there are $x/30$ integers divisible by 5 and 6. Subtracting all of them results in $x/4 + x/5 + x/6 - x/20 - x/12 - x/30$.

However, we also need to consider integers like 60, which are divisible by *all three* numbers 4, 5, 6, see Fig. 8.10b. While adding $x/4 + x/5 + x/6$, we have counted such integers three times. However, when subtracting $-x/20 - x/12 - x/30$, we have subtracted them three times, hence we have not counted them at all! Thus, we now need to add them back. There are $x/60$ such integers, hence the final formula should be

$$x/4 + x/5 + x/6 - x/20 - x/12 - x/30 + x/60 = 13x/30.$$

This method is called the *inclusion-exclusion* principle. If x is not divisible by 60, but is large, there are still approximately $x/4$ integers divisible by 4, about $x/5$ integers divisible by 5, and so on, and the same method tells us that about $13x/30$ integers are divisible by either 4, 5, or 6.

Integers Having a Divisor Between y and z

In general, for any x, y, z , let $H(x, y, z)$ be the number of integers $n \leq x$ having a divisor d such that $y < d \leq z$. We have just proved that $H(x, 2, 4) \approx x/2$ and $H(x, 3, 6) \approx 13x/30$ for large x . Let

$$\varepsilon(y, z) = \lim_{x \rightarrow \infty} \frac{H(x, y, z)}{x}.$$

Intuitively, $\varepsilon(y, z)$ represents the “density”, or “fraction” of all integers having a divisor between y and z . We have proved that $\varepsilon(2, 4) = 1/2$, while $\varepsilon(3, 6) = 13/30$. In 1935, Erdős [131] proved that $\lim_{y \rightarrow \infty} \varepsilon(y, 2y) = 0$, hence, for large y , there are very few integers having a divisor between y and $2y$.

Integers Having Exactly One Divisor Between y and z

Next, let us count the number of integers $n \leq x$ which have *exactly one* divisor from the set $\{3, 4\}$. For $x = 24$, these are $3, 4, 6, 8, 9, 15, 16, 18, 20, 21$, and the answer is 10. In general, there are $x/3$ integers divisible by 3, $x/4$ integers divisible by 4, and $x/12$ integers divisible by both. While adding $x/3 + x/4$, we count the last family twice, but now we should not count them at all, hence we need to subtract them twice, resulting in the formula $x/3 + x/4 - 2(x/12) = 5x/12$. In general, let $H_1(x, y, z)$ be the number of integers $n \leq x$ having *exactly one* divisor d such that $y < d \leq z$, and let

$$\varepsilon_1(y, z) = \lim_{x \rightarrow \infty} \frac{H_1(x, y, z)}{x}.$$

We have just proved that $\varepsilon_1(2, 4) = 5/12$.

Similarly, for $H_1(x, 3, 6)$, we add numbers $x/4, x/5$, and $x/6$ of integers divisible by 4, 5, and 6, respectively, then subtract $2(x/20)$ (twice the numbers divisible by 4 and 5), and, similarly, $2(x/12)$ and $2(x/30)$. Now, $x/60$ numbers divisible by 4, 5, 6 have been counted three times but subtracted six times, hence we need to add them three times back, resulting in

$$H_1(x, 3, 6) \approx x/4 + x/5 + x/6 - 2(x/20 + x/12 + x/30) + 3(x/60) = x/3,$$

hence $\varepsilon_1(3, 6) = 1/3$.

Erdős' Conjecture and Its Generalizations

Note that $\frac{\varepsilon_1(2, 4)}{\varepsilon(2, 4)} = \frac{5/12}{1/2} = \frac{5}{6} \approx 0.83$, while $\frac{\varepsilon_1(3, 6)}{\varepsilon(3, 6)} = \frac{1/3}{13/30} = \frac{10}{13} \approx 0.77 < 0.83$. In 1960, Erdős [132] conjectured that the ratio $\frac{\varepsilon_1(y, 2y)}{\varepsilon(y, 2y)}$ becomes smaller and smaller as y grows, and in fact

$$\lim_{y \rightarrow \infty} \frac{\varepsilon_1(y, 2y)}{\varepsilon(y, 2y)} = 0.$$

In other words, the conjecture predicts that there should be much fewer integers having *exactly* one divisor between y and $2y$ than integers having *at least* one such divisor.

In the following years, people studied various generalizations of this conjecture, but could not resolve it one way or another. A possible generalization is, for any $r \geq 1$, study the quantity $H_r(x, y, z)$, the number of integers $n \leq x$ having *exactly* r divisors d such that $y < d \leq z$, and the corresponding limit

$$\varepsilon_r(y, z) = \lim_{x \rightarrow \infty} \frac{H_r(x, y, z)}{x}.$$

We can then ask whether $\lim_{y \rightarrow \infty} \frac{\varepsilon_r(y, 2y)}{\varepsilon(y, 2y)} = 0$.

Even more generally, while Erdős' conjecture studies the case $z = 2y$, it would be natural to ask the same question for $z = \lambda y$ for $\lambda \neq 2$, or, for example, for $z = y^{1.01}$, etc.

The Solution

All these questions and generalizations have been answered by the following theorem of Kevin Ford [154].

Theorem 8.10 *For every $\lambda > 1$ and $r \geq 1$, there is a constant $\delta > 0$, depending on λ and r , such that*

$$\frac{\varepsilon_r(y, \lambda y)}{\varepsilon(y, \lambda y)} \geq \delta.$$

On the other hand, for each $r \geq 1$,

$$\lim_{z/y \rightarrow \infty} \frac{\varepsilon_r(y, z)}{\varepsilon(y, z)} = 0.$$

As a very special case ($\lambda = 2, r = 1$), Theorem 8.10 proves that Erdős' conjecture is false: the ratio $\frac{\varepsilon_1(y, 2y)}{\varepsilon(y, 2y)}$ is bounded below by some positive constant δ and therefore cannot converge to 0. The same is true if $z = \lambda y$ for any fixed λ . However, if z grows faster than a linear function on y , for example, $z = y^{1.01}$, then $z/y \rightarrow \infty$, and the ratio $\frac{\varepsilon_1(y, 2y)}{\varepsilon(y, z)}$ indeed goes to 0.

On the other hand, Theorem 8.10 is just one of many corollaries of a much more general result of Ford, which determines the order of magnitude of $H(x, y, z)$ for *all* values of x, y, z , not necessarily in the limit $x \rightarrow \infty$. It also determines the order of magnitude of $H_r(x, y, z)$ for a large region of values r, x, y, z .

An Application to Erdős' Multiplication Table Problem²

Consider the ordinary multiplication table, familiar from grade school. How many numbers appear in the table? If the table is 10×10 , then the answer is 100, right? But wait, some numbers appear more than once, such as 10, which appears four times ($1 \times 10, 2 \times 5, 5 \times 2, 10 \times 1$). If we interpret the question as how many *different* numbers appear as products in the table, then the answer for a 10×10 table is 42. If the table is enlarged to $N \times N$, then how many distinct products will appear? This is a problem posed by Paul Erdős [134] in 1955. For small values of N , this can be solved with a computer, but what happens for really large N ? Is there a formula, or a good approximation?

There probably isn't a "simple" formula for $M(N)$, the number of distinct products in an $N \times N$ multiplication table, but it is possible to obtain an approximate formula. The problem is closely related to counting numbers with a divisor in a given interval. If we consider, for example, a product in the multiplication table that is close to N^2 , then both factors are less than N but their product is close to N^2 , hence both factors are close to N . Thus the product is a number with a divisor very close to its square-root. Making this sort of argument precise, we arrive at the inequalities

$$H\left(\frac{N^2}{4}, \frac{N}{4}, \frac{N}{2}\right) \leq M(N) \leq \sum_{k=0}^{\infty} H\left(\frac{N^2}{2^k}, \frac{N}{2^{k+1}}, \frac{N}{2^k}\right).$$

The upper bound is deduced by observing that one of the factors must lie in an interval $(N/2^{k+1}, N/2^k]$ for some k and then the product is at most $N^2/2^k$. For the lower bound, if $n \leq N^2/4$ has a divisor $d \in (N/4, N/2]$, then $n = dk$, where $k = n/d \leq N$. Using the formulas for $H(x, y, z)$ proved in [154], it follows that

$$c_1 \frac{N^2}{(\ln N)^E (\ln \ln N)^{3/2}} \leq M(N) \leq c_2 \frac{N^2}{(\ln N)^E (\ln \ln N)^{3/2}},$$

where c_1, c_2 are positive constants, and $E = 1 - \frac{1+\ln \ln 2}{\ln 2} = 0.086071332\dots$. In particular, as was proved by Erdős, $\lim_{N \rightarrow \infty} \frac{M(N)}{N^2} = 0$; that is, most of the numbers between 1 and N^2 do not appear in the table! The above formula provides a precise measure of just how few elements do appear.

Reference

K. Ford, The distribution of integers with a divisor in a given interval, *Annals of Mathematics* **168**-2, (2008), 367–433.

²This section was written by Kevin Ford.

8.11 An Upper Bound for the Norm of the Inverse of a Random Matrix

Linear Transformations of Lines and Planes

Any function $f : \mathbb{R} \rightarrow \mathbb{R}$ can be viewed as a transformation of the real line \mathbb{R} : every point $x \in \mathbb{R}$ moves to $f(x)$. Perhaps the simplest transformations correspond to linear functions f with $f(0) = 0$, that is, functions of the form $f(x) = ax$ for some $a \in \mathbb{R}$. We will write such a function as “ $x \rightarrow ax$ ”. Geometrically, this corresponds to stretching the real line if $|a| > 1$, and a contraction if $|a| < 1$, together with reflection with respect to origin if $a < 0$. In the special case $a = 0$, the whole real line moves to the origin.

Similarly, any transformation of the coordinate plane can be viewed as a function $A : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, that is, a function sending any pair of real numbers (x, y) to another pair (u, v) . Again, the simplest examples are linear functions sending $(0, 0)$ to itself, which are functions of the form $(x, y) \rightarrow (ax + by, cx + dy)$ for some real coefficients a, b, c, d . Easy examples are (a) the “identity function” $(x, y) \rightarrow (x, y)$, sending each point to itself, (b) reflections $(x, y) \rightarrow (-x, -y)$ and $(x, y) \rightarrow (x, -y)$ with respect to the coordinate center and the x -axis, respectively, (c) the rotation $(x, y) \rightarrow (-y, x)$ by the angle 90° counter-clockwise, (d) the contraction $(x, y) \rightarrow (0.5x, 0.5y)$ of the plane with coefficient 0.5, (e) the projection $(x, y) \rightarrow (x, 0)$ to X axes, etc., see Fig. 8.11. A slightly more complicated example is $(x, y) \rightarrow (3x, 0.5y)$, stretching the plane in one direction and contracting it in another.

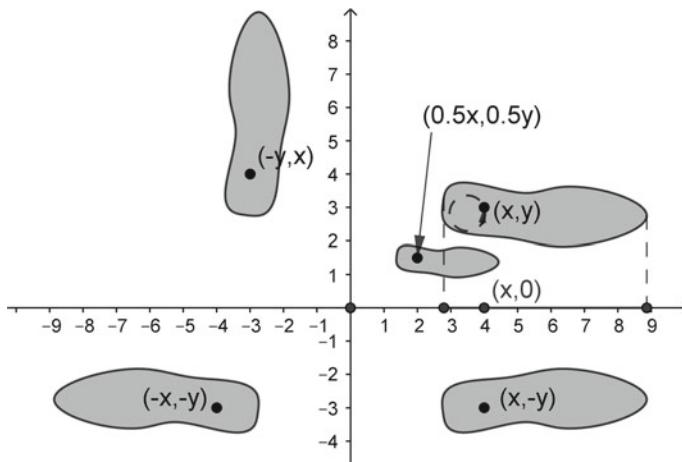


Fig. 8.11 Some linear transformations of the coordinate plane

Maximal Stretching and Maximal Contraction

A fundamental property of any such transformation A is its “maximal stretching and maximal contraction”. That is, let $z = (x, y)$ be any non-zero vector in the plane with length $|z| = \sqrt{x^2 + y^2} > 0$, let $Az = (ax + by, cx + dy)$ be its image after transformation A , and let $|Az| = \sqrt{(ax + by)^2 + (cx + dy)^2}$ be the length of the image. Then the norm of A , denoted $\|A\|$, is the maximal possible ratio $\frac{|Az|}{|z|}$ over all possible non-zero vectors z . Intuitively, this corresponds to the “maximal stretching”. For example, if A is given by $(x, y) \rightarrow (3x, 0.5y)$, and $z = (1, 2)$, then $Az = (3, 1)$, and $\frac{|Az|}{|z|} = \sqrt{2}$, while for $z = (1, 0)$ we have $Az = (3, 0)$, and $\frac{|Az|}{|z|} = 3$. We will show below that this is the maximal possible, hence $\|A\| = 3$. Similarly, we denote by $\|A^{-1}\|$ the maximal possible ratio $\frac{|z|}{|Az|}$, which corresponds to the maximal contraction. Then $1/\|A^{-1}\|$ is the minimal possible value of $\frac{|Az|}{|z|}$. In our example with A being $(x, y) \rightarrow (3x, 0.5y)$,

$$\frac{|Az|}{|z|} = \frac{\sqrt{9x^2 + 0.25y^2}}{\sqrt{x^2 + y^2}} = \sqrt{9u + 0.25(1-u)} = \sqrt{8.75u + 0.25},$$

where $u = \frac{x^2}{x^2+y^2}$. Because $0 \leq u \leq 1$, $\frac{|Az|}{|z|}$ is maximal and minimal when $u = 1$ and $u = 0$, respectively. This implies that $\|A\| = \sqrt{8.75 \cdot 1 + 0.25} = 3$ and $\|A^{-1}\| = 1/\sqrt{8.75 \cdot 0 + 0.25} = 2$. For the projection $(x, y) \rightarrow (x, 0)$, there are non-zero vectors z , for example, $z = (1, 0)$, such that $Az = (0, 0)$, and $|Az| = 0$. In this case we can treat $\frac{|z|}{|Az|}$ as infinity, and say that $\|A^{-1}\| = +\infty$.

Stretching and Contraction for Random Transformations

A fundamental question asks what the values of $\|A\|$ and $\|A^{-1}\|$ are for a *random* transformation A . By this, we mean that the coefficients a, b, c, d in the formula $(x, y) \rightarrow (ax + by, cx + dy)$ are selected at random. For example, we can toss a fair coin and assign $a = 1$ if we have got a head, and $a = -1$ if we have got a tail, and then repeat this experiment for b, c , and d . If, for example, it happens that $a = 1$, $b = -1$, $c = 1$, and $d = 1$, then A is the transformation $(x, y) \rightarrow (x - y, x + y)$, and

$$\frac{|Az|}{|z|} = \frac{\sqrt{(x-y)^2 + (x+y)^2}}{\sqrt{x^2 + y^2}} = \sqrt{2},$$

hence $\|A\| = 1/\|A^{-1}\| = \sqrt{2}$. In a similar way, we can study all 16 possible outcomes of this random experiment, and calculate $\|A\|$ and $\|A^{-1}\|$ in each case. For $\|A\|$, we would get $\sqrt{2}$ in 8 cases out of 16, and 0 in the other 8 cases. We conclude that $\|A\|$ may be $\sqrt{2}$ with probability $\frac{8}{16} = 0.5$, and 0 with probability 0.5. Similarly, for $\|A^{-1}\|$, the possible answers are $1/\sqrt{2}$ or $+\infty$ with probabilities 0.5 each.

n-Variable Transformations

Similarly, one may study n -variable transformations A of the form $(x_1, x_2, \dots, x_n) \rightarrow (y_1, y_2, \dots, y_n)$. Once again, we focus on the linear ones, in which case

$$y_j = a_{1j}x_1 + a_{2j}x_2 + \dots + a_{nj}x_n, \quad j = 1, 2, \dots, n. \quad (8.11)$$

In total, we need n^2 coefficients a_{ij} , $i = 1, 2, \dots, n$, $j = 1, 2, \dots, n$, to describe such a transformation A . Usually, these coefficients are written as an $n \times n$ table, and A called an $n \times n$ matrix.

As before, let us choose each a_{ij} to be either 1 or -1 with equal chance (we need n^2 coin tosses for this), and the question is to estimate $\|A\|$ and $\|A^{-1}\|$. It turns out that upper bounds for $\|A\|$ are easy to derive, but upper bounds for $\|A^{-1}\|$ are very difficult, and this is the subject of the theorem described below.

Subgaussian Random Variables

More generally, we may assume that the coefficients a_{ij} in (8.11) are either -1 , or 0 , or 1 , each variant with probability $1/3$. Alternatively, they can be arbitrary real numbers from the interval $[-1, 1]$ selected uniformly at random, etc. In general, real numbers selected at random according to some rule are called *random variables*. For a random variable X , its expectation, denoted $E[X]$, is, intuitively, its average value. For example, if X assumes values $-1, 0$, or 1 with equal chance, then $E[X] = \frac{1}{3}(-1 + 0 + 1) = 0$. If X is the result of throwing a die, that is, a number from 1 to 6 with equal chances, then $E[X] = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$ (we call this random variable X_D). If X is chosen uniformly at random from some interval $[a, b]$, then $E[X] = \frac{b-a}{2}$, etc. An important characteristic of any random variable X is how close it is to its expectation, that is, how big the (absolute value of the) difference $X - E[X]$ is. The average of the square of this difference, $E[(X - E[X])^2]$, is called the *variance* of X and denoted $\text{Var}[X]$. For example, for X_D , $E[X] = 3.5$, $(X - 3.5)^2$ takes 6 values $(1 - 3.5)^2, (2 - 3.5)^2, \dots, (6 - 3.5)^2$, whose average is 19.5, hence $\text{Var}[X_D] = 19.5$.

We say that a random variable X has *bounded support* if $|X| \leq t$ with probability 1 for some constant $t > 0$. All the random variables we have considered so far satisfy this property, for example, X_D has bounded support with $t = 6$. However, the random variable X' taking values $1, 2, 3, 4, \dots$ with probabilities $1/2, 1/4, 1/8, 1/16, \dots$ does not have bounded support: it can take arbitrarily large values, although with tiny probabilities. A random variable X is called *subgaussian* if there exist constants C and c such that

$$\mathbb{P}(|X| > t) \leq Ce^{-ct^2}, \quad \forall t > 0,$$

where \mathbb{P} denotes the probability, and e is the base of the natural logarithm. It is easy to check that all random variables with bounded support are automatically subgaussian, but not vice versa: the random variable X' described above is subgaussian as well.

The Theorem

The following theorem of Rudelson [329] establishes an upper bound for $\|A^{-1}\|$ for subgaussian random models.

Theorem 8.11 *Let X be a subgaussian random variable with $E[X] = 0$ and $\text{Var}[X] = 1$. Let A be an $n \times n$ matrix whose coefficients a_{ij} are selected independently at random according to rule X . Then there are absolute constants C and c such that for any $\varepsilon > c/\sqrt{n}$, the inequality*

$$\|A^{-1}\| \leq C \cdot n^{3/2}/\varepsilon$$

holds with probability greater than $1 - \varepsilon$, if n is large enough.

For example, taking $\varepsilon = 2c/\sqrt{n}$, we get the bound $\|A^{-1}\| \leq \frac{C}{2c} \cdot n^2$, holding with probability $1 - 2c/\sqrt{n}$.

Reference

M. Rudelson, Invertibility of random matrices: norm of the inverse, *Annals of Mathematics* **168**-2, (2008), 575–600.

8.12 An Isoperimetric Inequality with Optimal Exponent

Optimal and Almost-Optimal Square-Perimeter to Area Ratio

One of the oldest problems in Mathematics is to determine the shape of the region in the plane with a given area which has minimal perimeter, or, equivalently, the shape of the region of given perimeter with maximal area. For example, when ancient people built their cities, they would aim for maximal area (for more people to have a space to live inside) and minimal perimeter (to minimize the length of the fence around the city, which would also make it easier to protect themselves against enemies). For example, a square city of area $V > 0$ has side length \sqrt{V} , hence its perimeter is $P = 4\sqrt{V}$. At the same time, a circle of area V has radius r such that $\pi r^2 = V$, hence $r = \sqrt{V/\pi}$, and the perimeter is $P = 2\pi r = 2\sqrt{\pi V} \approx 3.54\sqrt{V} < 4\sqrt{V}$. It turns out that the circular form is indeed optimal, no region of area V has perimeter smaller than $2\sqrt{\pi V}$. In other words,

$$\frac{P^2}{V} \geq 4\pi, \tag{8.12}$$

and in fact the circle is the only shape for which equality holds.

Now, imagine that some region R in the plane has perimeter $P(R)$ and area $V(R)$ such that $\frac{P(R)^2}{V(R)} = 4\pi + 0.000001$. How can this be? An obvious possibility is that R has almost circular form, with some very little modifications. Intuitively, one would guess that this is *the only* possibility, it seems unlikely that some completely different shape has almost optimal $\frac{P(R)^2}{V(R)}$ ratio to within 0.000001. However, can we rigorously prove this? That is, if $\frac{P(R)^2}{V(R)}$ is close to 4π , can we prove that the region R is “close” to being a circle.

Measuring Closeness to Being a Circle

It is non-trivial even to formulate such a result rigorously. What exactly is meant for a region R to be “close to being a circle”? Well, for a *given* circle B , we can look at the area of the set $R \Delta B$, called the *symmetric difference* of R and B . This is, by definition, the union of all points of R which are not in B and all points of B which are not in R , see Fig. 8.12a. The larger the area $V(R \Delta B)$, the further away R is from the circle B .

Of course, even if $V(R \Delta B)$ is less than, say, 10^{-6} , this does not mean that R has almost circular form. It may be that R is, say, a square, but a very small one, and B is a very small circle with the same center. What really matters is the area of $V(R \Delta B)$ *in comparison with* the area $V(B) = \pi r(B)^2$ of the circle B , where $r(B)$ is the radius of B . Ignoring the π factor, we may look at the ratio $\frac{V(R \Delta B)}{r(B)^2}$. If it is small, then R is indeed a small modification of the circle B , hence it has an “almost circular” form. If it is large, then... well, then R is not close to this particular circle B , but it may be close to a different circle, with different center, see Fig. 8.12b–c.

Now, for a given region R , let \mathcal{B} be the set of *all* circles with the same area as R , that is, with radius $r = \sqrt{V(R)/\pi}$ and all possible centres. Then what we need is

$$\lambda(R) = \min_{B \in \mathcal{B}} \frac{V(R \Delta B)}{r(B)^2}.$$

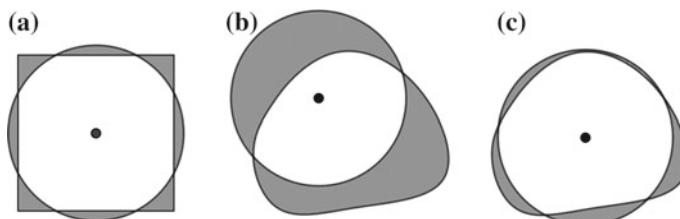


Fig. 8.12 Measuring closeness to being a circle

$\lambda(R)$ is called the *Fraenkel asymmetry* of R , and is indeed an indicator of how close R is to being a circle.

Isoperimetric Deficit and Rigorous Problem Formulation

For any region R , we can define its *isoperimetric deficit* as

$$D(R) = \frac{P(R) - P(B)}{P(B)},$$

where B is the circle having the same area as R (so that $P(B) = 2\sqrt{\pi V(R)}$). It measures how close the perimeter of R is from being optimal. Now, our question can finally be rigorously formulated as “find an upper bound on $\lambda(R)$ as a function of $D(R)$ ”.

The same question can be asked in higher dimensions. In three-dimensional space, a ball B with volume $V(B)$ has radius $r(B) = \sqrt[3]{\frac{3V(B)}{4\pi}}$, and surface area $P(B) = 4\pi r(B)^2 = \sqrt[3]{36\pi V(B)^2}$, which is again the smallest possible. As in the planar case, one may ask whether “approximately best” surface area $P(R)$ implies that the three-dimensional body R is “close” to being a ball. As before, the (relative) closeness of $P(R)$ to the optimal surface area $P(B)$ is measured by isoperimetric deficit $D(R)$, while its “closeness” to being a ball is measured by Fraenkel asymmetry $\lambda(R)$, with $r(B)^3$ instead of $r(B)^2$ in the denominator. The same definition also extends to any dimension n , with

$$\lambda(R) = \min_{B \in \mathcal{B}} \frac{V(R \Delta B)}{r(B)^n},$$

where \mathcal{B} is the set of all n -dimensional balls with $V(B) = V(R)$. Then, again, the problem is to find an upper bound on $\lambda(R)$ as a function of $D(R)$.

The Proof of Hall’s Conjecture

Hall [188] proved in 1992 that, for every dimension n , there is a constant C_n , such that

$$\lambda(R) \leq C_n \sqrt[4]{D(R)}$$

for all n -dimensional bodies R . Based on this, we are able to conclude that “if $D(R)$ is very small then so is $\lambda(R)$ ”, as requested. However, $\sqrt[4]{D(R)}$ converges to 0 rather slowly. For example, to conclude that $\lambda(R)/C_n \leq 0.01$, we need to have $D(R) \leq (0.01)^4 = 0.0000001$. Hence, Hall’s result becomes practically useful only if the perimeter is *extremely* close to being optimal. On the other hand, experiments with different shapes suggested that $D(R) \leq (\lambda(R)/C_n)^2$ (which is 0.0001 in our

example) should suffice for the same conclusion. This was the basis for Hall's conjecture, which had been open for 16 years, until its positive resolution in 2008.

Theorem 8.12 ([159]) *Let $n \geq 2$. There exists a constant C_n such that for every Borel set R in \mathbb{R}^n with $0 < V(R) < \infty$,*

$$\lambda(R) \leq C_n \sqrt[2]{D(R)}. \quad (8.13)$$

Here, \mathbb{R}^n denotes n -dimensional space (for example, \mathbb{R}^2 is the plane and \mathbb{R}^3 is the usual space we live in), and a "Borel set" is, intuitively, a set for which area/volume is well-defined (one can construct some exotic sets for which area or volume cannot be even defined, and, obviously, Theorem 8.12 does not have any meaning for such sets).

Can the Bound in Theorem 8.12 Be Improved?

The exponent 1/2 in (8.13) is the best possible, in the sense that the term $\sqrt[2]{D(R)}$ in the right-hand side of (8.13) cannot be replaced by $D(R)^\alpha$ for some $\alpha > 1/2$. In fact, in any dimension n , one can take the sequence of ellipsoids E_k

$$E_k = \{x_1^2(1 + \varepsilon_k) + \frac{x_2^2}{1 + \varepsilon} + x_3^2 + \dots + x_n^2 < 1\}$$

with $\varepsilon_k \rightarrow 0$, and check that $D(E_k) \rightarrow 0$ and that

$$\frac{1}{c} D(E_k) \leq \lambda(E_k)^2 \leq c D(E_k)$$

for some positive constant c .

However, to obtain the tightest possible bound, we should try to prove (8.13) with as small a constant C_n as possible. Theorem 8.12 only says that a constant C_n exists, but does not provide any explicit value for it. In a subsequent paper, Figalli, Maggi, and Pratelli [152] proved that (8.13) holds with an explicit constant $C_n = 181n^7/(2 - 2^{(n-1)/n})^{3/2}$, which gives, for example, $C_2 \approx 5 \cdot 10^4$ and $C_3 \approx 10^6$. However, these values of the constants C_n are definitely non-optimal, and the question of finding sharp values of C_n is open and looks very hard. Another interesting question, also seemingly very hard, is estimating the growth of C_n at the limit as n goes to infinity.

Reference

N. Fusco, F. Maggi, and A. Pratelli, The sharp quantitative isoperimetric inequality, *Annals of Mathematics* **168**-3, (2008), 941–980.

8.13 A Negative Answer to Maharam's Question

The Banach–Tarski Paradox and Sets with No Volume or Length

One of the most amazing facts in mathematics is the Banach–Tarski paradox. One can divide the 3-dimensional ball into 5 pieces, and then move these pieces around and rotate them to construct... two balls of the same radius! At first glance, this is impossible for obvious reason: if $V > 0$ is the volume of the ball, and V_1, V_2, \dots, V_5 are the volumes of the pieces, then $\sum_{i=1}^5 V_i = V$, and not $\sum_{i=1}^5 V_i = 2V$, hence constructing two balls from the same pieces is impossible. However, it turns out that the pieces are so complicated that the volume for them... just cannot be defined at all, so that this argument does not work.

There is no Banach–Tarski paradox in dimensions 1 and 2, in particular, one cannot divide the interval $[0, 1]$ into a finite number of pieces, and then translate them in such a way to fully cover the interval $[0, 2]$. However, even in dimension 1, there are some complicated subsets of $[0, 1]$ for which one cannot define a “length” in the usual sense.

The Vitali Set

The most well-known example of such a set is the Vitali set [394]. This is a subset $V \subset [0, 1]$ such that for any real number r , there is exactly one number $v \in V$ such that $v - r$ is a rational number (that is, can be written in the form $\frac{m}{n}$ for some integers $m, n, n \neq 0$). Why does such a set exist? Well, for any real number r , like $r = \sqrt{2}$, let S_r be the set of all real numbers $x \in [0, 1]$ such that $x - r$ is a rational number. For example, $S_{\sqrt{2}}$ contains numbers $x = \sqrt{2} - 1, x = \sqrt{2} - 1/2$, etc. Then all real numbers $x \in [0, 1]$ are divided into groups (called *equivalence classes*) such that each group is S_r for some r , see Fig. 8.13. Now, let us select one arbitrary element from each group (there is an axiom of mathematics, called the *axiom of choice*, which says that we can do this), and let V be the set of selected elements. Then V is the Vitali set.

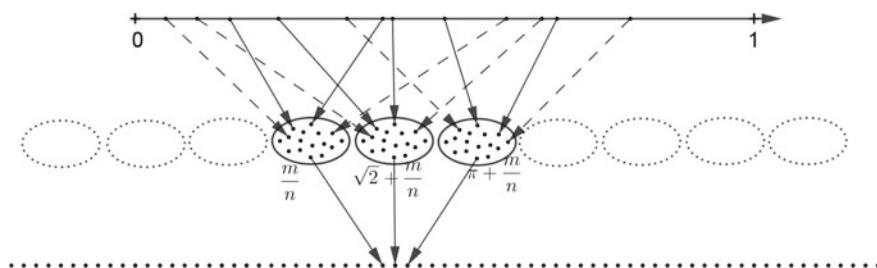


Fig. 8.13 The construction of the Vitali set

Why Can the Vitali Set Not Have a Length?

It is known that there exists a sequence x_1, x_2, x_3, \dots containing each rational number in $[-1, 1]$ exactly once. For example, such a sequence may start from

$$0, 1, -1, 1/2, -1/2, 2/3, 1/3, -1/3, -2/3, \dots$$

and then list all rational numbers in $[-1, 1]$ (not listed so far) with denominators 4, 5, 6, and so on. Now, let $V_k, k = 1, 2, 3, \dots$ be the set of all numbers representable in the form $v + x_k$ for some $v \in V$. In fact, each V_k is just a shifted copy of V , hence, if $l(V)$ is the length of V , then $l(V_k) = l(V), \forall k$.

It follows from the definition of V that all sets V_k are disjoint. Hence, if V^* is the union of all V_k , then

$$l(V^*) = \sum_{k=1}^{\infty} l(V_k) = \sum_{k=1}^{\infty} l(V).$$

In particular, if $l(V) = 0$ then $l(V^*) = 0$, while if $l(V) > 0$, then $l(V^*) = \infty$. Hence, in any case, $l(V^*)$ must be either 0 or ∞ . However, by definition, each element of V^* can be represented as $v + x_k$, where $0 \leq v \leq 1$ (because $v \in V$), and $-1 \leq x_k \leq 1$. Hence, $V^* \subset [-1, 2]$, and $l(V^*) \leq 3 < \infty$. On the other hand, by definition of V , every real number $x \in [0, 1]$ can be represented as $x = v + x_k$ where $v \in V$ and x_k is rational, hence $[0, 1] \subset V^*$, and $l(V^*) \geq 1 > 0$, a contradiction. This proves that $l(V)$ is not well-defined.

Finitely Additive Measures

If we cannot measure the lengths of all subsets of $[0, 1]$, what can we do? Well, we can generalize the notion of “length”. For example, in the above proof that the Vitali set does not have a length, we say that “because each V_k is just a shifted copy of V , $l(V_k) = l(V)$ ”. However, we can define the “length” of any interval $(a, b) \subset (0, 1)$ as, for example, $l'(a, b) = \int_a^b x dx$. In this case $l'(0, 1/2) = 1/8$, while $l'(1/2, 1) = 3/8 \neq 1/8$, despite the fact that the interval $(1/2, 1)$ is just the interval $(0, 1/2)$, shifted by $1/2$.

Also, we have used that $l(V^*) = \sum_{k=1}^{\infty} l(V_k)$, provided that all V_k are disjoint. It is possible to prove the existence of some “length” l^* , which is defined for *all* subsets of $[0, 1]$, such that

$$l^*(A \cup B) = l^*(A) + l^*(B), \quad \text{whenever } A \text{ and } B \text{ are disjoint,}$$

and hence $l^*(\bigcup_{k=1}^m V_k) = \sum_{k=1}^m l^*(V_k)$, but this equality fails if $m = \infty$. Such an l^* is called a *finitely additive measure*, and its existence suffices to show that the Banach–Tarski paradox is impossible in dimension 1. However, any such l^* is difficult to explicitly describe.

Outer Measure

As another option, we can first consider only some family \mathcal{F} of “nice” subsets for which length is definable. For example, we can consider only subsets which are countable unions of intervals, that is, let \mathcal{F} be the family of subsets X of $[0, 1]$ of the form

$$X = (a_1, b_1) \cup (a_2, b_2) \cup \cdots \cup (a_k, b_k) \cup \dots \quad (8.14)$$

where the endpoints a_k and b_k of each interval can be included or not. Then the length $l(X)$ is well defined and is given by $l(X) = \sum_{k=1}^{\infty} (b_k - a_k)$.

A simple and intuitive attempt to define the “length” of all subsets of $[0, 1]$ would be the following: for any $S \subset [0, 1]$, let \mathcal{F}_S be the collection of all X of the form (8.14) such that $S \subset X$, and let $l(S)$ be the largest real number such that $l(S) \leq l(X)$, $\forall X \in \mathcal{F}_S$. Then $l(S)$ is called the *outer measure*, it is defined for all subsets of $[0, 1]$, and

$$l(A \cup B) \leq l(A) + l(B), \quad \text{for all } A, B \subset [0, 1].$$

However, we may have $l(A \cup B) < l(A) + l(B)$, even if A and B are disjoint.

Measures and Submeasures

Formally, a collection \mathcal{F} of subsets of some set U is called a *Boolean algebra of sets*, if $A \cup B \in \mathcal{F}$, $A \cap B \in \mathcal{F}$, and $A^c \in \mathcal{F}$, whenever $A, B \in \mathcal{F}$, where A^c is the complement of set A , that is, $A^c = \{x \in U \mid x \notin A\}$. A map $\mu : \mathcal{F} \rightarrow [0, \infty)$ is called a (finitely additive) *measure* if (i) $\mu(\emptyset) = 0$, where \emptyset denotes the empty set, (ii) $\mu(A) \leq \mu(B)$ if $A, B \in \mathcal{F}$ and $A \subset B$, and (iii) $\mu(A \cup B) = \mu(A) + \mu(B)$, whenever A, B are disjoint. For example, subsets of $[0, 1]$ of the form (8.14) form a Boolean algebra of sets, and the “usual” length $l(X) = \sum_{k=1}^{\infty} (b_k - a_k)$, $X \in \mathcal{F}$, is a measure. A map $\mu : \mathcal{F} \rightarrow [0, \infty)$ is called a *submeasure* if it satisfies (i), (ii), and (iv) $\mu(A \cup B) \leq \mu(A) + \mu(B)$, $A, B \in \mathcal{F}$. Every measure is a submeasure, but not vice versa: the outer measure on subsets of $[0, 1]$ is an example of a submeasure which is not a measure.

Exhaustive Submeasures

In fact, μ such that $\mu(A) = 1$, $\forall A \neq \emptyset$, is another (and completely uninteresting) example of a submeasure which is not a measure. Such uninteresting examples can be excluded if we study only *exhaustive* submeasures. A submeasure μ is called *exhaustive* if $\lim_{n \rightarrow \infty} \mu(E_n) = 0$ whenever $E_1, E_2, \dots, E_n, \dots$ are all disjoint. This seems to be a quite a natural property. In particular, any measure is exhaustive. Indeed, if μ is a measure, and $\mu(E_n)$ does not converge to 0, then there exist an $\varepsilon > 0$, and an infinite number of indices $n_1, n_2, \dots, n_k, \dots$ such that $\mu(E_{n_k}) \geq \varepsilon$ for all k . Then

(iii) implies that $\mu(E_{n_1} \cup E_{n_2} \cup \dots \cup E_{n_k}) \geq k\varepsilon$, $\forall k$. Because everything is a subset of U , then, by (ii), $k\varepsilon \leq \mu(U)$, $\forall k$, which is a contradiction if $k > \frac{\mu(U)}{\varepsilon}$.

Moreover, any submeasure ν which is absolutely continuous with respect to some measure μ must be exhaustive. ν is called *absolutely continuous* with respect to μ if for any $\varepsilon > 0$ there is a $\delta > 0$ such that $\mu(A) \leq \delta$ implies $\nu(A) \leq \varepsilon$, for all $A \in \mathcal{F}$. If ν is not exhaustive, then essentially the same argument as above leads to a contradiction.

A Negative Answer to Maharam's Question

Maharam's famous problem [257], formulated in 1947, asks if the converse is true, that is, if *every* exhaustive submeasure must be absolutely continuous with respect to some measure. The following theorem of Talagrand [368] proves that, somewhat unexpectedly, the answer is negative.

Theorem 8.13 *There exists a Boolean algebra \mathcal{F} of sets, and a nonzero exhaustive submeasure ν on it, which is not absolutely continuous with respect to any measure.*

Reference

M. Talagrand, Maharam's problem, *Annals of Mathematics* **168**-3, (2008), 981–1009.

Chapter 9

Theorems of 2009



9.1 On De Giorgi's Conjecture in Dimension at Most 8

The Derivative and Differential Equations

One of the central concepts in the whole of mathematics is the *derivative*, which for a function $u : \mathbb{R} \rightarrow \mathbb{R}$ is defined as $u'(x) = \lim_{\varepsilon \rightarrow 0} (u(x + \varepsilon) - u(x))/\varepsilon$, provided that the limit exists. There are numerous useful rules which help us to calculate derivatives for various functions, for example $(u + v)'(x) = u'(x) + v'(x)$, $(uv)'(x) = u'(x)v(x) + u(x)v'(x)$, $u'(x) = f'(g(x))g'(x)$ if $u = f(g(x))$, etc. We also have tables of derivatives for various functions, e.g. $(x^n)' = nx^{n-1}$ for any $n \neq 0$, $(\ln x)' = \frac{1}{x}$, $x > 0$, etc.

If we know the derivative of a function, we can recover it (up to an additive constant) using integration: if $u'(x) = f(x)$ then $u(x) = \int f(x)dx + C$ for some $C \in \mathbb{R}$. For example, $u'(x) = 2x$ implies that $u(x) = \int 2x dx + C = x^2 + C$. An equation of the form $u'(x) = f(x)$ is the simplest example of a *differential equation*, that is, an equation involving derivatives. Differential equations arise in numerous application of mathematics, like physics, biology, and economics, and they are extremely important in pure mathematics as well.

A Slightly More Complicated Differential Equation

A slightly more complicated differential equation is of the form $u'(x) = f(u)$, for example

$$u'(x) = 1 - u^2. \quad (9.1)$$

To solve such an equation, we use the notation $\frac{du}{dx}$ for the derivative, rewrite equation $\frac{du}{dx} = f(u)$ in the form $\frac{du}{f(u)} = dx$, and take the integral of both sides. For example, Eq. (9.1) can be written as $\frac{du}{1-u^2} = dx$, or $\int \frac{du}{1-u^2} = \int dx$. A table of integrals tells

us that $\int \frac{du}{1-u^2} = -\frac{1}{2} \ln \frac{1-u}{1+u} + C_1$ for $|u| < 1$, and $\int dx = x + C_2$, so the equation reduces to $\ln \frac{1-u}{1+u} = -2(x + C)$, where $C = C_2 - C_1$, or $\frac{1-u}{1+u} = e^{-2(x+C)}$, where e is the base of the natural logarithm. This results in $u(x) = \tanh(x + C)$, $C \in \mathbb{R}$, where \tanh is the function defined by $\tanh(x) := \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

Differential Equations Involving the Second Derivative

A differential equation may also include the second derivative $u''(x) = (u'(x))'$. For example, if $u(x) = x^3$ then $u'(x) = 3x^2$, and $u''(x) = (3x^2)' = 6x$. Another notation for the second derivative is $\frac{d^2u}{dx^2}$, so the differential equation $\frac{d^2u}{dx^2} = 6x$ has a solution $u(x) = x^3$ (and many other solutions as well).

How could we find at least one non-trivial solution (the trivial solution is $u(x) = 0, \forall x$) of a more complicated equation involving the second derivative, like

$$\frac{d^2u}{dx^2} = u^3 - u \quad ? \quad (9.2)$$

Because the second derivative is a polynomial in u , one could naturally guess that the first derivative u' might also be written as $P(u)$ for some polynomial P . If it were a linear polynomial, that is, $u' = Au + B$ for some $A, B \in \mathbb{R}$, then $u'' = (Au + B)' = Au' = A(Au + B)$, again a linear polynomial, not a cubic one as in (9.2).

Next, let us try a quadratic polynomial P , that is, assume that $u' = Au^2 + Bu + C$, where $A, B, C \in \mathbb{R}$ are some unknown coefficients to be found from (9.2)—this is called the *method of undetermined coefficients*. The second derivative is $u'' = (Au^2 + Bu + C)' = A(u^2)' + Bu' = A(2uu') + Bu' = (2Au + B)u' = (2Au + B)(Au^2 + Bu + C) = (2A^2)u^3 + (3AB)u^2 + (2AC + B^2)u + BC$. By (9.2), this expression should be equal to $u^3 - u$, which leads us to the system of equations $2A^2 = 1$, $3AB = 0$, $2AC + B^2 = -1$, $BC = 0$, which has 2 solutions, one of which is $A = -1/\sqrt{2}$, $B = 0$, $C = 1/\sqrt{2}$, in which case $u' = Au^2 + Bu + C = (1/\sqrt{2})(1 - u^2)$.

The last equation, up to the constant factor $1/\sqrt{2}$, coincides with (9.1), and, using exactly the same argument as above, we get a family of solutions

$$u(x) = \tanh \left(\frac{x + C}{\sqrt{2}} \right), \quad C \in \mathbb{R}. \quad (9.3)$$

Note that each function from family (9.3) satisfies two additional conditions: (i) $|u(x)| < 1$ for every $x \in \mathbb{R}$ (this follows from the definition of $\tanh(x)$) and (ii) $u'(x) > 0$ for every $x \in \mathbb{R}$ (this follows from $u'(x) = 1 - u^2$ and (i)). This implies that the functions (9.3) are increasing, see Fig. 9.1.

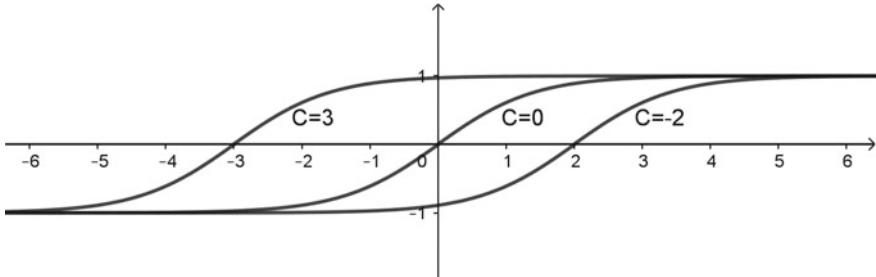


Fig. 9.1 Some solutions (9.3) to Eq.(9.2)

Differential Equations with Functions of Two Variables

Derivatives and differential equations can also be studied for functions of *several* variables, such as $u = u(x, y)$. The *partial* derivative $\frac{\partial u}{\partial x}$ for such a function is the derivative with respect to x if y is treated as a constant. The derivative $\frac{\partial u}{\partial y}$ with respect to y is defined similarly, the second-order derivative $\frac{\partial^2 u}{\partial x^2}$ is $\frac{\partial}{\partial x} \left(\frac{\partial u}{\partial x} \right)$, etc. For example, if $u = x^3 y$, then $\frac{\partial u}{\partial y} = x^3$, $\frac{\partial u}{\partial x} = 3x^2 y$, $\frac{\partial^2 u}{\partial x^2} = 6xy$, and so on. Would you be able to find at least one non-trivial solution of an equation involving second partial derivatives, like

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = u^3 - u \quad ? \quad (9.4)$$

One of the simplest ideas is to try to combine x and y in a linear way, so that $u(x, y) = g(Ax + By + C)$ for some $A, B, C \in \mathbb{R}$ and function $g : \mathbb{R} \rightarrow \mathbb{R}$. In this case, $\frac{\partial u}{\partial x} = g'(Ax + By + C) \cdot \frac{\partial(Ax + By + C)}{\partial x} = Ag'(Ax + By + C)$, and therefore $\frac{\partial^2 u}{\partial x^2} = Ag''(Ax + By + C) \cdot \frac{\partial(Ax + By + C)}{\partial x} = A^2 g''(Ax + By + C)$. Similarly, $\frac{\partial^2 u}{\partial y^2} = B^2 g''(Ax + By + C)$. Substituting this back into the equation and defining $z = Ax + By + C$, we get $(A^2 + B^2)g''(z) = g^3(z) - g(z)$. If $A^2 + B^2 = 1$, this is exactly the Eq. (9.2) above, which has solutions in the form $g(z) = \tanh\left(\frac{z+C}{\sqrt{2}}\right)$. Hence, for every $C \in \mathbb{R}$, and every A and B such that $A^2 + B^2 = 1$, the function $u(x, y) = \tanh\left(\frac{Ax+By+C}{\sqrt{2}}\right)$ is a solution to (9.4). Note that if we choose $B > 0$, this solution satisfies the condition (i) $|u(x, y)| < 1$ and (ii) $\frac{\partial u}{\partial y} > 0$ for every $x, y \in \mathbb{R}$, and also has a special geometric structure: if we fix any $\lambda \in (-1, 1)$, the set of points (x, y) such that $u(x, y) = \lambda$ (such sets are called the *level sets*) forms a line $Ax + By + C = z$, where z is such that $\tanh(z) = \lambda$.

Differential Equations with Functions of n Variables

We can write an equation similar to (9.4) for functions $u = u(x_1, x_2, \dots, x_n)$ in *any* number of variables, that is,

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2} = u^3 - u \quad (9.5)$$

and ask to find solutions satisfying the same conditions (i) $|u| < 1$ and (ii) $\frac{\partial u}{\partial x_n} > 0$ for every $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. In fact, this particular differential equation originates in the theory of phase transition, and is very important and well-studied. By a similar argument as above we may write a solution

$$u(x) = \tanh\left(\frac{(x - p) \cdot b}{\sqrt{2}}\right), \quad (9.6)$$

where $p = (p_1, p_2, \dots, p_n) \in \mathbb{R}^n$, $b = (b_1, b_2, \dots, b_n) \in \mathbb{R}^n$ is such that $\sum_{i=1}^n b_i^2 = 1$ and $b_n > 0$, and \cdot denotes the scalar product in \mathbb{R}^n (that is $(x - p) \cdot b = \sum_{i=1}^n (x_i - p_i)b_i$). Again, for any fixed λ , the level set of points $x \in \mathbb{R}^n$ such that $u(x) = \lambda$ is described by a linear equation of the form $(x - p) \cdot b = z$ and is called a *hyperplane*. However, guessing one family of solutions to (9.5) does not mean finding *all* the solutions.

The De Giorgi Conjecture

In 1978, De Giorgi [111] made a conjecture that, for $n \leq 8$, *every* solution to (9.5) satisfying (i) and (ii) on the whole of \mathbb{R}^n has the property that all level sets are hyperplanes, and therefore is given by (9.6). By 2009, the conjecture has been proved only for $n = 2, 3$. The following theorem, proved in [333], is a big advance for $4 \leq n \leq 8$.

Theorem 9.1 Suppose u is a solution to (9.5) such that (i) $|u| < 1$; (ii) $\frac{\partial u}{\partial x_n} > 0$ for every $x = (x_1, \dots, x_n) \in \mathbb{R}^n$; and (iii) $\lim_{x_n \rightarrow \pm\infty} u(x_1, \dots, x_{n-1}, x_n) = \pm 1$ for every fixed x_1, \dots, x_{n-1} . If $n \leq 8$, then all level sets of u are hyperplanes.

In other words, Theorem 9.1 proves the De Giorgi conjecture for all $n \leq 8$, under the additional condition (iii). In a later work, Manuel del Pino, Michal Kowalczyk and Juncheng Wei [113] found a counterexample to the De Giorgi conjecture in all dimensions $n \geq 9$, hence the condition $n \leq 8$ in Theorem 9.1 cannot be removed.

Reference

O. Savin, Regularity of flat level sets in phase transitions. *Annals of Mathematics* **169**-1, (2009), 41–78.

9.2 An Efficient Algorithm for Fitting a Smooth Function to Data

Looking for a “Nice” Function Which Fits the Given Data Set

In Sect. 5.2, we discussed the following question. Assume that we have performed some measurements, like the radiation level along a street, at a finite number of points. Can we then “guess” the result of the measurement at any other point?

Mathematically, let $F(\cdot)$ be the (unknown) function such that $F(x)$ represents the level of radiation at any point x . Let x_1, x_2, \dots, x_N be the points at which we have performed the measurements, and y_1, y_2, \dots, y_N be the corresponding results. Then we need to find a function $F(x)$ such that $F(x_i) = y_i$, $i = 1, \dots, N$, see Fig. 9.2a.

Of course, there are infinitely many ways of doing this, for example, we can put $F(x_i) = y_i$, $i = 1, \dots, N$ and $F(x) = 0$ for all other x . However, it is unlikely that the true function F is anything like this. At the very least, we would expect it to be continuous and “smooth”. So, the true question is to find F such that (a) $F(x_i) = y_i$, $i = 1, \dots, N$, and (b) F is a function which is as “nice” as possible.

Which Functions are “Nice”?

However, how do we define which functions are “nice” and which are not, and moreover, measure this “niceness” numerically? One standard approach is to require that

- (i) The function F does not take too large or too small values;
- (ii) The function F is smooth and does not increase or decrease too fast. Mathematically, this means that the derivative F' exists and does not take too large or too small values;
- (iii) In turn, the rate of increase/decrease of F does not change too suddenly. This means that F' does not change its values too fast, or, equivalently, that the second derivative of F , denoted $F^{(2)}$, does not take too large or too small values;
- (iv) And so on.

Based on this intuition, one natural way to measure the “niceness” of a function $F : \mathbb{R} \rightarrow \mathbb{R}$ is its C^m norm, defined as

$$\|F\|_{C^m} := \max\{\sup_{x \in \mathbb{R}} |F(x)|, \sup_{x \in \mathbb{R}} |F'(x)|, \dots, \sup_{x \in \mathbb{R}} |F^{(m)}(x)|\},$$

where $F^{(m)}$ is the m -th derivative of F . Our problem can then be formalized as

$$\min_F \|F\|_{C^m}, \quad \text{s.t. } F(x_i) = y_i, \quad i = 1, \dots, N.$$

“Nice” Functions of Several Variables

If we measure the radiation in a city, not along a street, then $x_i \in \mathbb{R}^2$, $i = 1, \dots, N$, are points in the plane, and the aim is to find the “nicest” function $F : \mathbb{R}^2 \rightarrow \mathbb{R}$ such that $F(x_i) = y_i$, $i = 1, \dots, N$. For measurements in space, $x_i \in \mathbb{R}^3$, $i = 1, \dots, N$, and $F : \mathbb{R}^3 \rightarrow \mathbb{R}$. More generally, the result of a measurement can depend on n parameters, and, in this case, each x_i , $i = 1, 2, \dots, N$ is a point in \mathbb{R}^n , and the aim is to find the “nicest” function $F : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $F(x_i) = y_i$, $i = 1, \dots, N$.

The “niceness” can be defined similarly as above, because the definition of the C^m norm can be extended to functions $F : \mathbb{R}^n \rightarrow \mathbb{R}$. If $F(z_1, z_2, \dots, z_n)$ is such a function, we can assume that the variables $z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n$ are fixed, and treat F as a function g of one variable z_i . The derivative of g is called the *partial derivative* of F with respect to z_i . This derivative can then be differentiated again with respect to some other variable, and so on. Let us assume that we are allowed to perform m such differentiations, and then evaluate the resulting derivative at any point we wish. For example, let $n = m = 3$, and the function $F(x, y, z) = xy^2z^3$. We can first differentiate it with respect to, say, z , to get $3xy^2z^2$, then with respect to x to get $3y^2z^2$, and then with respect to z again to get $6y^2z$. Finally, we can substitute any values, say, $x = 1$, $y = 2$, $z = 3$, to get a numerical value $6 \cdot 2^2 \cdot 3 = 72$. For any function $F : \mathbb{R}^n \rightarrow \mathbb{R}$, its C^m norm $\|F\|_{C^m(\mathbb{R}^n)}$ is the maximal possible absolute value of the number which we can get in this way after up to m differentiations. We can then minimize such a norm subject to $F(x_i) = y_i$, $i = 1, \dots, N$.

Approximate Fitting to the Dataset

In fact, we can also assume that the measurements can be done with some error, and relax the requirement $F(x_i) = y_i$, $i = 1, \dots, N$ to $|F(x_i) - y_i| \leq M\sigma_i$, $i = 1, \dots, N$, see Fig. 9.2b. Here, σ_i , $i = 1, \dots, N$ are some given non-negative numbers, and M is a variable which should be made as small as possible—the smaller M , the better the function F approximates our data y_i .

Finally, our problem becomes

$$\min_F M, \quad \text{s.t.} \quad \|F\|_{C^m(\mathbb{R}^n)} \leq M \quad \text{and} \quad |F(x_i) - y_i| \leq M\sigma_i, \quad i = 1, \dots, N. \quad (9.7)$$

Can We Solve Problem (9.7) Efficiently?

In applications, the dimension n and parameter m are fixed, but N can be very large. Can we solve the optimization problem (9.7) efficiently? A theorem of Fefferman, which we discussed in Sect. 5.2, implies the existence of an algorithm solving (9.7) in time proportional to N^k , where k is a large constant, which depends on n and

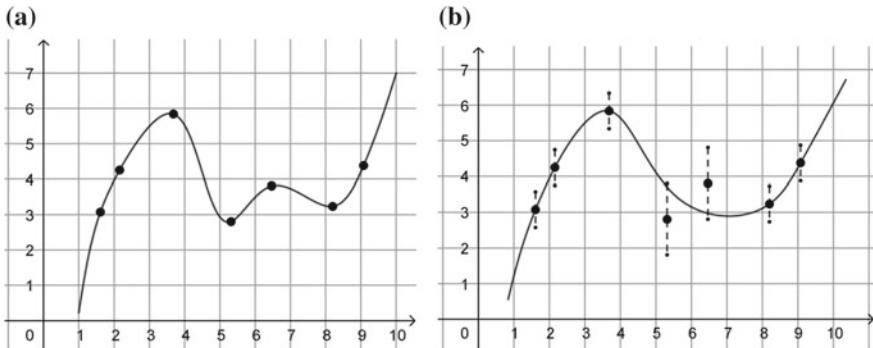


Fig. 9.2 Exact and approximate fitting of a function to data

m . Of course, even for $k = 10$ or $k = 15$, the running time N^k becomes impractical already for $N = 100$, while in applications N can be measured in millions.

It is very difficult, and most probably impossible, to solve problem (9.7) both efficiently and exactly. For practical purposes, an approximate solution is often sufficient. We say that an algorithm computes *the order of magnitude* of the solution to (9.7) if it returns output M' such that $aM' \leq M^* \leq bM'$, where M^* is the optimal solution of (9.7), and a and b are some constants, depending only on n and m .

An Efficient Algorithm for the Order of Magnitude

The following theorem of Fefferman and Klartag [146] states that the order of magnitude of the solution to (9.7) can be computed very efficiently.

Theorem 9.2 *There exists an algorithm which computes the order of magnitude of the optimal value in (9.7) using at most $CN \ln N$ operations and at most CN memory, where C is a constant which depends only on m and n .*

In Theorem 9.2, “operations” means the usual operations with real numbers, such as addition, subtraction, multiplication, division, or comparison. It is assumed that all these operations are performed with perfect accuracy. “ CN memory” means CN memory cells, each of which can store a real number, again with perfect accuracy. Of course, in reality any irrational real number has an infinite amount of digits, hence we need to round such numbers to store them, and all operations are subject to rounding errors. However, these issues are minor and were addressed in subsequent publications by the authors.

Also, Theorem 9.2 just computes (the order of magnitude of) *the optimal value* in the optimization problem (9.7). Of course, what we really need is to construct a function F for which this optimal value is achieved. In subsequent work [147], the authors developed an algorithm for this task as well.

At first, it is not clear how an algorithm can return a function. After all, a function is defined by its values at every point, and there are infinitely many points. In fact, Fefferman and Klartag's algorithm works in two stages. At stage one, it takes input data $(m, n, x_i, y_i, \sigma_i)$ and does some preprocessing. At stage 2, it takes any point $x_0 \in \mathbb{R}^n$ as an input, and returns a polynomial which approximates F in a neighbourhood of the point x_0 .

Reference

C. Fefferman and B. Klartag, Fitting a C^m -smooth function to data I, *Annals of Mathematics* **169**-1, (2009), 315–346.

9.3 A Helicoid-Like Surface with a Hole

How “Curved” is a Curve?

Given a curve, how do we measure how “curved” it is? For this, the concept of *curvature* is used. Intuitively, the curvature of any curve at any point is just the “speed of rotation” at this point, while you are travelling along the curve at unit speed. As a simple example, imagine you are travelling along a circle of radius R with unit speed. Then it is clear that your “speed of rotation” is the same throughout your journey. Because it takes you time $2\pi R$ to rotate around the full angle of 360° , or 2π radians, your “speed of rotation” per unit of time is $\frac{2\pi}{2\pi R} = \frac{1}{R}$. In other words, the curvature of a circle is the same at every point and is equal to $\frac{1}{R}$. As another simple example, when you are travelling along a straight line, there is no rotation at all, hence the curvature is 0.

The Curvature of a Parabola

In general, of course, a curve may be straight, or almost straight, at some places, but be “curved a lot” elsewhere, so its curvature may vary from point to point. For example, let us estimate the curvature of the parabola $y = x^2$ near some point $x = x_0$. After some short time, you would travel from point (x_0, x_0^2) to point (x, x^2) , where $x = x_0 + \varepsilon$ for some small ε . Then $x^2 = (x_0 + \varepsilon)^2 = x_0^2 + 2x_0\varepsilon + \varepsilon^2 \approx x_0^2 + 2x_0\varepsilon = 2x_0x - x_0^2$. Hence, the direction of your movement is along the line $y = 2x_0x - x_0^2$, which is parallel to $y = 2x_0x$. In other words, you move with angle α with respect the x -axis, so that $\tan \alpha = 2x_0$.

By the same logic, at the final point $(x_0 + \varepsilon, (x_0 + \varepsilon)^2)$ your angle of movement β is such that $\tan \beta = 2(x_0 + \varepsilon)$. By the trigonometric formula, $\tan(\beta - \alpha) = \frac{\tan \beta - \tan \alpha}{1 + \tan \beta \tan \alpha} = \frac{2\varepsilon}{1 + 2(x_0 + \varepsilon) \cdot 2x_0} \approx \frac{2\varepsilon}{1 + 4x_0^2}$. Because $\beta - \alpha$ is small, $\beta - \alpha \approx \tan(\beta - \alpha) \approx \frac{2\varepsilon}{1 + 4x_0^2}$. This is how much you have rotated yourself.

How much time do you need for this? The distance you have travelled is

$$\sqrt{(x - x_0)^2 + (x^2 - x_0^2)^2} = \sqrt{(x - x_0)^2 + (x - x_0)^2(x + x_0)^2} \approx \varepsilon \sqrt{1 + 4x_0^2}.$$

Because your speed was 1, you needed time $\varepsilon \sqrt{1 + 4x_0^2}$. Hence, your rotation speed is $\frac{2\varepsilon}{1+4x_0^2} : (\varepsilon \sqrt{1 + 4x_0^2}) = \frac{2}{(1+4x_0^2)^{3/2}}$. Note that the curvature is maximal at $x_0 = 0$ and is almost 0 for large x_0 , which agrees well with our visual impression that the parabola is most “curved” at 0 and looks almost like a straight line “further away” from 0, see Fig. 9.3a.

If we travel along the curve $y = x^2$ from left to right, the direction of travel rotates counter-clockwise. If we travel along the curve $y = -x^2$, the rotation at $x = x_0$ has the same magnitude $\left(\frac{2}{(1+4x_0^2)^{3/2}}\right)$ but opposite direction, clockwise, and, to emphasize this fact, we can say that in this case the curvature is negative and is equal to $-\frac{2}{(1+4x_0^2)^{3/2}}$.

The Curvature of Any Curve

In general, the direction of movement at the point $x = x_0$ along a curve $y = f(x)$ is parallel to the line $y = bx$, where $b = f'(x_0)$ is the derivative of f at x_0 , see Sect. 3.1. A calculation very similar to the one above suggests that the “speed of rotation” is given by the formula

$$k = \frac{f''(x_0)}{(1 + (f'(x_0))^2)^{3/2}}, \quad (9.8)$$

where $f''(x_0)$ denotes the second derivative of f (that is, the derivative of the function $f'(x)$). In particular, for $f(x) = x^2$ we have $f'(x) = 2x$ and $f''(x) = 2$, so that $k = \frac{2}{(1+4x_0^2)^{3/2}}$, confirming the calculation above. In fact, we can now forget about the initial semi-formal discussion and just use Eq. (9.8) as the definition of the curvature of any curve which is the graph of a twice differentiable function f . For example, the *catenary curve* is defined by the equation

$$y = a \cosh\left(\frac{x}{a}\right) = \frac{a}{2}(e^{x/a} + e^{-x/a}),$$

where $\cosh(t) = \frac{1}{2}(e^t + e^{-t})$ is called the *hyperbolic cosine* of t , and $a > 0$ is the parameter, see Fig. 9.3b for some examples of graphs of catenary curves. The catenary has a physical interpretation as “the curve that an idealized cable assumes under its own weight when supported only at its ends”. Substitution of $f(x) = a \cosh\left(\frac{x}{a}\right)$ into (9.8) yields the curvature $k = \frac{1}{a \cosh(x/a)} = \frac{1}{f(x)}$.

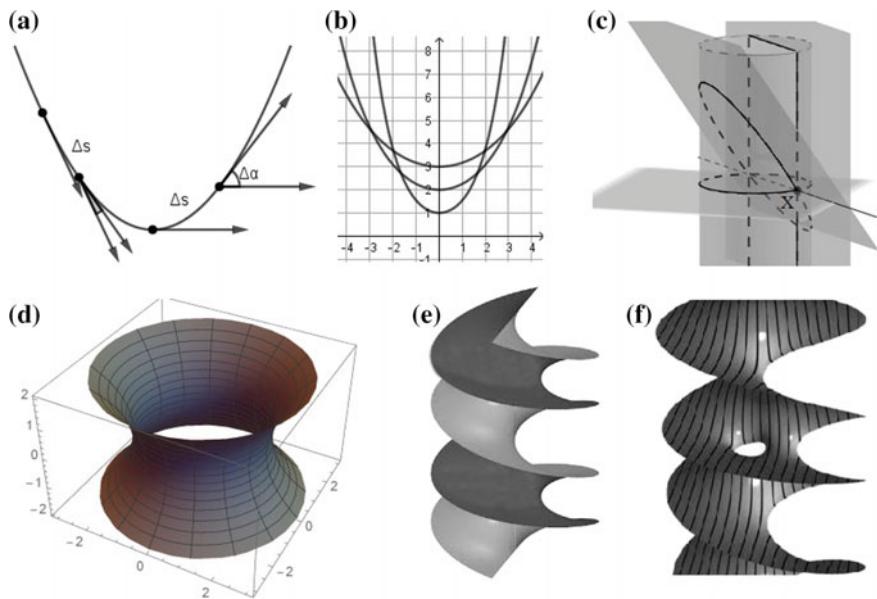


Fig. 9.3 **a** Curvature of a parabola; **b** Catenary curves; **c** Principal curvatures of a cylinder; **d** Catenoid; **e** Helicoid; **f** Helicoid with a hole

How “Curved” Is a Surface?

How do we measure “how curved” a two-dimensional surface S is in \mathbb{R}^3 ? Well, at any point X of the surface S , we can build a vector perpendicular to it, choose a plane containing this vector (called a “normal plane”), and measure the curvature at X of the curve which is the intersection of the surface and the plane. For example, if S is a sphere with radius R , then the intersection of any normal plane with S is just a circle of radius R , and its curvature is $1/R$.

In general, however, the answer depends on the choice of normal plane. For example, if S is an infinite cylinder with base circle having radius R , and S is any point on S , then one normal plane intersects S on a circle with radius R and curvature $1/R$, while another one intersects S on two parallel lines with curvature 0. We can also construct “intermediate” normal planes intersecting the cylinder in an ellipse, see Fig. 9.3c, and the curvature at X would be between 0 and $1/R$.

Principal, Mean, Gaussian, and Total Curvatures

In general, the minimal and maximal curvatures at X over all choices of normal planes are denoted k_1 and k_2 and called the *principal curvatures* of S at X . Their mean $H = (k_1 + k_2)/2$ is called the *mean curvature*, while their product $K = k_1 k_2$ is called the *Gaussian curvature* of S at X . The integral of the Gaussian curvature

over the whole surface is called the *total curvature* of S . For example, if S is a sphere of radius R , the principal curvatures are $k_1 = k_2 = 1/R$, the mean curvature is also $1/R$, the Gaussian curvature is $1/R^2$, and the total curvature is (Gaussian curvature)·(Surface volume) = $(1/R^2)(4\pi R^2) = 4\pi$. If S is a cylinder, the principal curvatures are 0 and $1/R$, the mean curvature is $1/2R$, the Gaussian curvature is 0, and hence the total curvature is 0 as well.

To obtain a cylinder, we can take a line $y = R$ in the x - y coordinate plane, and rotate it in three-dimensional space around the x -axis. If, instead of a line, we rotate a catenary curve $y = a \cosh(x/a)$, the corresponding surface is called a *catenoid*, see Fig. 9.3d. For a point $X = (x, a \cosh(x/a), 0)$ on the catenoid S , one normal plane is the x - y coordinate plane, which intersects S at a catenary curve, whose curvature at X is $\frac{1}{a \cosh(x/a)}$. One can show that this is the maximal possible, and the minimal possible is $-\frac{1}{a \cosh(x/a)}$. Hence, in this case, the principal curvatures are $\pm \frac{1}{a \cosh(x/a)}$, hence the mean curvature is identically 0. A surface with mean curvature identically 0 is called a *minimal surface*, see Sect. 5.3 for an alternative definition of this concept and a detailed discussion. A trivial example of a minimal surface is the plane, while the catenoid is the first non-trivial example, found by Euler in 1744. The Gaussian curvature of the catenoid is $-\frac{1}{a^2 \cosh^2(x/a)}$, and its total curvature turns out to be -4π .

Properly Embedded Curves and Surfaces

A plane curve is a set of points (x, y) in the Euclidean plane \mathbb{R}^2 such that $x = x(t)$, $y = y(t)$, $t \in I$, where $x(t)$ and $y(t)$ are continuous functions, and I is some (finite or infinite) interval of the real numbers. An example of a curve is the parabola $x = t$, $y = t^2$, $t \in \mathbb{R}$. A curve is called *simple* if it has no self-intersection, that is, for all $t_1, t_2 \in I$, if $x(t_1) = x(t_2)$ and $y(t_1) = y(t_2)$ then $t_1 = t_2$. For example, the parabola is a simple curve, while the curve $x(t) = t^3 - t$, $y(t) = t^2$, $t \in \mathbb{R}$ is not simple, because $(x(-1), y(-1)) = (x(1), y(1)) = (0, 1)$. Another example of a simple curve is $x(t) = \frac{\sin t}{t}$, $y(t) = \frac{\cos t}{t}$, $0 < t < +\infty$, known as a *hyperbolic spiral*. If $t \rightarrow +\infty$, this curve winds around $(0, 0)$, approaches it, but never reaches it. The part of the curve corresponding to the infinite interval $1 \leq t < +\infty$ is contained in the bounded closed unit disk $\{(x, y) : x^2 + y^2 \leq 1\}$.

A curve $C \subset \mathbb{R}^2$ (or a surface $S \subset \mathbb{R}^3$) is called *properly embedded* in \mathbb{R}^2 (respectively, \mathbb{R}^3), if it has no self-intersections, and its intersection with any compact subset of \mathbb{R}^2 (respectively, \mathbb{R}^3) is compact. Intuitively, this means that no “infinite” part of the curve or surface is contained in any finite region. For example, the parabola is properly embedded in \mathbb{R}^2 , while the hyperbolic spiral is not, because the ‘infinitely long’ part of the spiral is contained in a small region around $(0, 0)$. Planes and catenoids are examples of properly embedded surfaces in \mathbb{R}^3 .

The Helicoid and Its “Generalizations”

After the catenoid, the next discovered example of a properly embedded minimal surface in \mathbb{R}^3 was the *helicoid*. This is the surface given by

$$x = s \cos(\alpha t), \quad y = s \sin(\alpha t), \quad z = t,$$

where α is a constant, and s, t are real parameters, ranging from $-\infty$ to ∞ , see Fig. 9.3e, and also Sect. 5.3. Unlike planes and catenoids, the helicoid has *infinite* total curvature.

A surface S is said to have *finite topology* if it is homeomorphic to a compact surface with a finite number of points removed (that is, it can be obtained from such a surface via a continuous transformation, see Sect. 8.1 for more details). Since the proof that the helicoid is a minimal surface in 1776, a lot of minimal surfaces have been discovered, but none of them had finite topology and infinite total curvature, and it was an important open question whether such a surface exists, besides the helicoid. This question was resolved positively in 2009.

Theorem 9.3 ([398]) *There exists a properly embedded minimal surface in \mathbb{R}^3 with finite topology and infinite total curvature, which is not a helicoid.*

An example of a surface satisfying the conditions of Theorem 9.3 has got a name: the “embedded genus-one helicoid”. It looks like a helicoid with a hole, see Fig. 9.3f.

Reference

M. Weber, D. Hoffman and M. Wolf, An embedded genus-one helicoid, *Annals of Mathematics* **169**-2, (2009), 347–448.

9.4 Bounding the Condition Number of Random Discrete Matrices

Linear Functions of One and Two Variables

A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is called linear if $f(x + y) = f(x) + f(y)$ for all $x, y \in \mathbb{R}$. With $y = 0$, this implies $f(x + 0) = f(x) + f(0)$, hence $f(0) = 0$. With $y = -x$, we get $0 = f(0) = f(x + (-x)) = f(x) + f(-x) = 0$, hence $f(-x) = -f(x)$ for all x . Such functions are called *odd* functions.

With $y = x$, we get $f(2x) = f(x) + f(x) = 2f(x)$. Then $f(3x) = f(2x) + f(x) = 3f(x)$, and, by induction, $f(nx) = nf(x)$ for all x and all non-negative integers n . Because f is an odd function, this implies that in fact $f(nx) = nf(x)$, $\forall x \in \mathbb{R}$, for all integers n .

If $f(1) = a$, and $m, n \neq 0$ are any integers, then $f(m) = f(m \cdot 1) = m \cdot f(1) = ma$, hence $ma = f(m) = f\left(n \cdot \frac{m}{n}\right) = nf\left(\frac{m}{n}\right)$, hence $f\left(\frac{m}{n}\right) = \frac{m}{n}a$. In other words,

$f(x) = ax$ for all rational numbers x . If we also assume that f is continuous, this implies that $f(x) = ax$ for all $x \in \mathbb{R}$. For example, $f(x) = 2x$ and $f(x) = x/2$ are linear functions.

Similarly, a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, transforming a pair of real numbers (x, y) into another pair (u, v) , is called linear if $f(x_1 + x_2, y_1 + y_2) = f(x_1, y_1) + f(x_2, y_2)$. By an argument similar to the one above, one can prove that any linear continuous f has the form $f(x, y) = (ax + by, cx + dy)$ for some real coefficients a, b, c, d . For example, $f(x, y) = (x + y, -x - y)$ and $f(x, y) = (x + y, x - y)$ are linear functions.

Stretching and Contraction

A linear function f is called a stretching if $|f(x)| > |x|$ for all x , and a contraction if $|f(x)| < |x|$ for all x . In the one-variable case, $f(x) = ax$ is a stretching if $|a| > 1$ and a contraction if $|a| < 1$. For functions of two variables, the situation may be more involved. For example, the function $f(x, y) = (x + y, -x - y)$ is, geometrically, a composition of a projection, clockwise rotation, and a homothetic transformation with coefficient 2, see Fig. 9.4a, and it can stretch some vectors and contract others.

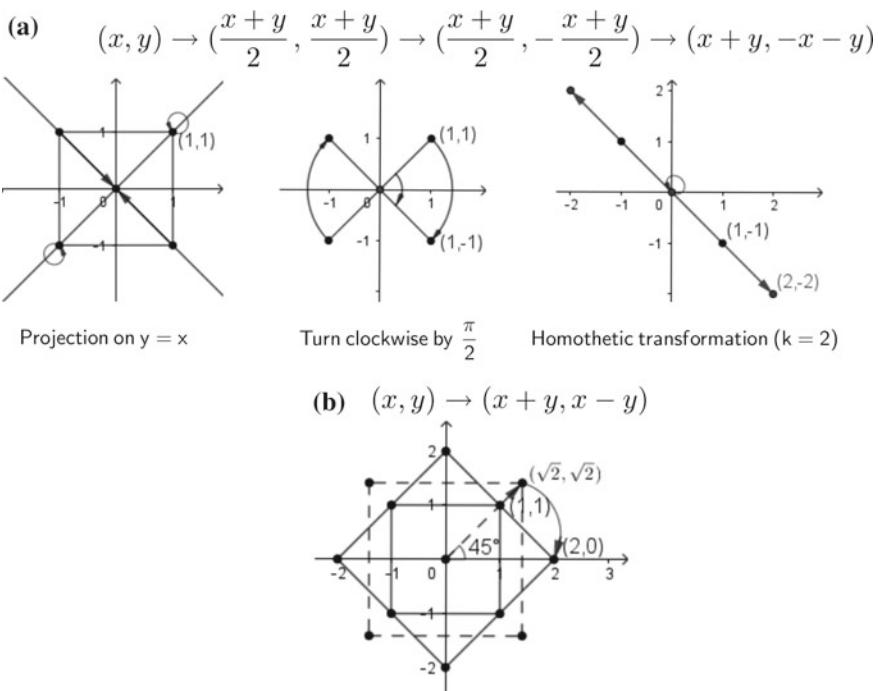


Fig. 9.4 Geometry of transformations **a** $f(x, y) = (x+y, -x-y)$, **b** $f(x, y) = (x+y, x-y)$

For example, it sends the vector $(1, 1)$ to $(1+1, -1-1) = (2, -2)$. The length of $(1, 1)$ is $\|(1, 1)\| = \sqrt{1^2 + 1^2} = \sqrt{2}$, while the length of $f(1, 1)$ is $|f(1, 1)| = \sqrt{2^2 + (-2)^2} = 2\sqrt{2}$, hence the vector $(1, 1)$ has been stretched twice. On the other hand, $f(1, -1) = (1-1, -1-(-1)) = (0, 0)$, that is, the non-zero vector $(1, 1)$ has been contracted to 0.

Let $\sigma(f)$ denote the minimal possible ratio of $\frac{|f(z)|}{|z|}$ over all $z = (x, y)$ with $|z| = \sqrt{x^2 + y^2} \neq 0$.

As we have seen above, $\sigma(f) = 0$ for $f(x, y) = (x+y, -x-y)$. On the other hand, for the function $f(x, y) = (x+y, x-y)$ we have

$$\frac{|f(z)|}{|z|} = \frac{\sqrt{(x+y)^2 + (x-y)^2}}{\sqrt{x^2 + y^2}} = \frac{\sqrt{2x^2 + 2y^2}}{\sqrt{x^2 + y^2}} = \sqrt{2},$$

hence $\sigma(f) = \sqrt{2}$. Geometrically, this function is a composition of a homothetic transformation (with $k = \sqrt{2}$) and a rotation, see Fig. 9.4b, and therefore stretches every vector in the same way.

The Role of $\sigma(f)$ in the Task of Inverting f

Given a function f and the value $f(x, y)$, can we uniquely determine x and y ? For the function $f(x, y) = (x+y, x-y)$, this is always possible. For example, if $f(x, y) = (3, -1)$, then $x+y = 3$ and $x-y = -1$, which can be easily solved to get $x = 1, y = 2$. In general, it is easy to prove that uniquely “restoring” (x, y) given $f(x, y)$ is always possible if $\sigma(f) > 0$. However, for a function with $\sigma(f) = 0$ this procedure does not work. For example, for the function $f(x, y) = (x+y, -x-y)$ assume that $f(x, y) = (3, -1)$. Then $x+y = 3$ and $-x-y = -1$. However, $x+y = 3$ implies $-x-y = -3 \neq -1$, a contradiction. On the other hand, if $f(x, y) = (1, -1)$, then $x+y = 1$ and $-x-y = -1$ which is possible for many different pairs x, y , for example, $x = 0$ and $y = 1$ or $x = 2$ and $y = -1$, etc.

For this reason, functions f with $\sigma(f) = 0$ are “unpleasant” in applications. Also, if $\sigma(f) > 0$ but is very small, then the “restoring” procedure above is possible, but may be difficult to compute. Hence, the ideal situation is when we can prove that $\sigma(f)$ is not 0, and moreover is “reasonably far away” from 0.

Estimating the Proportion of “Bad” Functions

How many “good” and “bad” functions are there? For simplicity, assume that the coefficients a, b, c, d in the formula $f(x, y) = (ax+by, cx+dy)$ can be either $+1$ or -1 . Because there are two options for each coefficient, there are $2^4 = 16$ such functions in total. An easy (but boring) computation shows that exactly 8 of them (including $f(x, y) = (x+y, -x-y)$ discussed above) have $\sigma(f) = 0$, and another 8 (including $f(x, y) = (x+y, x-y)$) have $\sigma(f) = \sqrt{2}$. In other words, if

we select such a function f at random, we get $\sigma(f) = 0$ with probability $\frac{8}{16} = 0.5$, and $\sigma(f) = \sqrt{2}$ with probability $\frac{8}{16} = 0.5$ as well.

A proportion of 50% of bad functions looks discouraging, but the situation improves if we consider more general functions $A_n : \mathbb{R}^n \rightarrow \mathbb{R}^n$, transforming n -tuples (x_1, x_2, \dots, x_n) into (y_1, y_2, \dots, y_n) . If such A_n is linear and continuous, then

$$y_j = a_{1j}x_1 + a_{2j}x_2 + \cdots + a_{nj}x_n, \quad j = 1, 2, \dots, n,$$

where a_{ij} , $i = 1, 2, \dots, n$, $j = 1, 2, \dots, n$ are real coefficients. It is standard to write the coefficients in an $n \times n$ table, with a_{ij} being at the intersection of the i -th row and j -th column, and then A_n is called an $n \times n$ *matrix*. Once again, assume that each a_{ij} is either $+1$ or -1 . Because there are n^2 coefficients, we have 2^{n^2} such functions/matrices A_n in total. A famous result of Kahn et al [213] states that the proportion of those matrices having $\sigma(A_n) = 0$ is at most 0.999^n . This result is not useful for small n (for $n = 2$, it gives the estimate $0.999^2 \approx 0.998$ while we know that the true proportion is 0.5), but, for large n , the expression 0.999^n decreases rapidly. For example, while the bound $0.999^{1,000} \approx 0.37$ is still not very useful, the bound $0.999^{10,000} \approx 0.00005$ for $n = 10,000$ is already good, while the bound $0.999^{100,000} \approx 3.5 \cdot 10^{-44}$ for $n = 100,000$ is much better than needed for any practical purposes. In fact, in later work [377] the bound 0.999^n has been improved to approximately 0.5^n , which already gives an excellent estimate $0.5^{30} \approx 9.3 \cdot 10^{-10}$ for $n = 30$.

How Often Is $\sigma(f)$ Small?

However, as mentioned above, a function/matrix A_n is practically unpleasant for the inversion procedure even if $\sigma(A_n)$ is positive but small. For concreteness, let us agree that by “small” we mean smaller than $\frac{1}{n^B} = n^{-B}$ for some constant $B > 0$. So, the problem is to find a good upper bound for the proportion of matrices A_n with $\sigma(A_n) < n^{-B}$, or, equivalently, for the probability $P(\sigma(A_n) < n^{-B})$ that a randomly selected A_n has small $\sigma(A_n)$.

A theorem of Rudelson [329], discussed in Sect. 8.11, implies that $P(\sigma(A_n) < C\varepsilon n^{-3/2}) \leq \varepsilon$ holds for sufficiently large n and for any $\varepsilon > c/\sqrt{n}$, where C and c are some constants. This is significant progress, but the condition $\varepsilon > c/\sqrt{n}$ is restrictive in some important applications. Even for large n like $n = 10,000$, $1/\sqrt{n}$ is 0.01, hence (assuming for simplicity that $c = 1$) the above theorem works only for $\varepsilon > 0.01$, and provides no more than 99% warranty that $\sigma(A_n)$ is small.

This was the state of the art before the following theorem was proved by Terence Tao and Van Vu [372].

Theorem 9.4 *For any positive constant A , there is a positive constant B such that for any sufficiently large n*

$$P(\sigma(A_n) < n^{-B}) \leq n^{-A}.$$

The importance of Theorem 9.4 is that it works for any $A > 0$. For example, selecting $A = 10$ we get an estimate n^{-10} for the proportion of “unpleasant” matrices A_n . For $n = 10,000$, this gives¹ a chance of just $10,000^{-10} = 10^{-40}$ for $\sigma(A_n)$ to be “small”, which is a much better probability guarantee than in Rudelson’s result.

Reference

T. Tao and V. Vu, Inverse Littlewood-Offord theorems and the condition number of random discrete matrices, *Annals of Mathematics* **169**-2, (2009), 595–632.

9.5 Characterizing the Legendre Transform of Convex Analysis

Convex Sets and Functions

A region S in the plane is called *convex* if it contains the straight line segment AB whenever points A and B belong to S , see Fig. 9.5a. For example, any disk or the area bounded by a triangle is a convex region. On the other hand, if ABC is a triangle, and X is any point strictly inside it, then the area bounded by the quadrilateral $ABXC$ is non-convex, because it contains points B and C but not the line segment BC , see Fig. 9.5b.

A more complicated example of a convex region is the set of all points (x, y) in the coordinate plane such that $y \geq x^2$. In general, for any function $f : \mathbb{R} \rightarrow \mathbb{R}$, the set of all points (x, y) such that $y \geq f(x)$ is called the *epigraph* of f , and a function f is called *convex* if its epigraph is a convex set. Equivalently, a function f is convex if

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \quad (9.9)$$

for all $x, y \in \mathbb{R}$ and all $\lambda \in [0, 1]$. For $f(x) = x^2$ this reduces to $(\lambda x + (1 - \lambda)y)^2 \leq \lambda x^2 + (1 - \lambda)y^2$, which simplifies to $\lambda(1 - \lambda)(x^2 - 2xy + y^2) \geq 0$, or equivalently $\lambda(1 - \lambda)(x - y)^2 \geq 0$.

For readers familiar with the concept of ‘derivative’ there is a much simpler proof that $f(x) = x^2$ is a convex function. The derivative of f is $f'(x) = 2x$, and the second derivative is $f''(x) = 2$. There is a theorem that if $f''(x)$ exists and is positive for all $x \in \mathbb{R}$, then f is a convex function.

Convex Sets as Intersections of Half-Planes

Another example of a convex region is the set of all points (x, y) in the coordinate plane such that $y \geq |x|$, where $|\cdot|$ denotes the absolute value. This is the epigraph

¹In fact, Theorem 9.4 works only for “sufficiently large” n , and there is no warranty that it works for $n = 10,000$. However, we think that this calculation is still useful for the purpose of illustration.

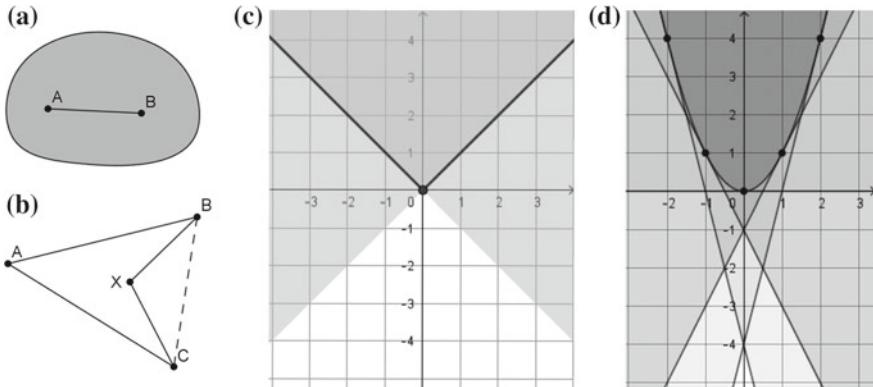


Fig. 9.5 **a** A convex set, **b** A non-convex set, **c** and **d** Convex sets as intersections of half-planes

of the convex function $f(x) = |x|$. This function is not differentiable at 0, but its convexity easily follows directly from (9.9).

The inequality $y \geq |x|$ can be equivalently written as “ $y \geq x$ and $y \geq -x$ ”. Geometrically, the set of points (x, y) satisfying an inequality of the form $y \geq ax + b$ for some constants a, b is a half-plane. Hence, the epigraph of the function $f(x) = |x|$ is the intersection of the half-planes $y \geq x$ and $y \geq -x$, see Fig. 9.5c. Representing a set as an intersection of half-planes is extremely convenient in optimization, where linear inequalities are the easiest to deal with.

Can the epigraph $S = \{(x, y) \mid y \geq x^2\}$ of the function $f(x) = x^2$, see Fig. 9.5d, be written as an intersection of half-planes? This looks unlikely, because the intersection of any finite number of half-planes has a piecewise linear boundary, while in our case the boundary is smooth. However, what if we allow an *infinite* number of half-planes? A half-plane $H(a, b) = \{(x, y) \mid y \geq ax + b\}$ contains S if $x^2 \geq ax + b$ for all x . For example, the half-plane $H(1, 0) = \{(x, y) \mid y \geq x\}$ does not contain S because the inequality $f(x) = x^2 \geq x$ does not hold for, say $x = 0.5$. On the other hand, $H(1, -1/4) = \{(x, y) \mid y \geq x\}$ contains S , because the inequality $x^2 \geq x - 1/4$ is equivalent to $(x - 1/2)^2 \geq 0$ and is valid for all x . In fact, $H(1, b)$ contains S if and only if $b \leq -1/4$. More generally, $H(a, b)$ contains S if and only if $b \leq -a^2/4$. Indeed, if $b \leq -a^2/4$, or $0 \leq -b - a^2/4$, then the inequality $x^2 \geq ax + b$ can be written as $x^2 - ax + a^2/4 - a^2/4 - b \geq 0$, or $(x - a/2)^2 + (-b - a^2/4) \geq 0$, which clearly holds for all x . On the other hand, if $b > -a^2/4$, the inequality $x^2 \geq ax + b$ fails for $x = -a/2$. In summary, the set $S = \{(x, y) \mid y \geq x^2\}$ is the intersection of half-planes $H(a, -a^2/4)$, and the inequality $y \geq x^2$ is equivalent to an infinite number of linear (in x) inequalities $y \geq ax + (-a^2/4)$, $a \in \mathbb{R}$.

The Legendre Transform and Its Properties

In general, the half-plane $H(a, b) = \{(x, y) \mid y \geq ax + b\}$ contains an epigraph $S = \{(x, y) \mid y \geq f(x)\}$ of a function $f(x)$ if and only if

$$f(x) \geq ax + b, \quad \forall x.$$

This holds if and only if $b \leq -(ax - f(x))$ for all x , or, equivalently, if and only if $b \leq -\phi(a)$, where the function $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is defined by

$$\phi(a) := \max_{x \in \mathbb{R}}(ax - f(x)).$$

If f is a convex function, S is the intersection of half-planes $H(a, -\phi(a))$, $a \in \mathbb{R}$, and the non-linear inequality $y \geq f(x)$ is equivalent to an infinite number of linear (in x) inequalities $y \geq ax - \phi(a)$, $a \in \mathbb{R}$.

The function ϕ is called the *Legendre transform* of the function f , and we write $\phi = Lf$. Unfortunately, the Legendre transform is not always well-defined as a finite-valued function. For example, if $f(x) = x$, and $a = 3$, then the expression $ax - f(x) = 3x - x = 2x$ can be arbitrarily large for large x . In this case, we put $\phi(3) = +\infty$. In general, we may allow our functions f and ϕ to take infinite values, and then the Legendre transform is always well-defined. Moreover, the Legendre transform Lf of any convex function f is always a convex function itself. In addition, the Legendre transform has some other useful properties, for example

- (P1) $LLf = f$ for any convex function f ;
- (P2) $f \leq g$ implies $Lf \geq Lg$.

In (P1), by LLf we mean “the Legendre transform of the Legendre transform of f ”, while in (P2) by $f \leq g$ we mean $f(x) \leq g(x)$ for all x . For example, we have proved above that the Legendre transform of the function $f(x) = x^2$ is the function $\phi(a) = a^2/4$. By absolutely the same argument, we can prove that the Legendre transform of $f(x) = Cx^2$ is $\phi(a) = a^2/4C$ for any constant C . With $C = 1/4$, this implies that the Legendre transform of $f(x) = x^2/4$ is $\phi(a) = a^2$, in agreement with (P1). With $C = 2$, this implies that the Legendre transform of $f(x) = 2x^2$ is $\phi(a) = a^2/8$. Note that we have $x^2 \leq 2x^2$ for all x , but $a^2/4 \geq a^2/8$ for all a , in agreement with (P2).

The Legendre Transform of Multivariate Convex Functions

The definition of convexity (9.9) works equally well for functions of *several* variables, for example, $f(x, y) = x^2 + y^2$, and, more generally, $f(x_1, x_2, \dots, x_n) = x_1^2 + x_2^2 + \dots + x_n^2$ are convex functions. A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex if and only if its epigraph $S = \{(x_1, \dots, x_n, y) \mid y \geq f(x_1, x_2, \dots, x_n)\}$ is a convex subset of \mathbb{R}^{n+1} . The inequality $y \geq f(x_1, x_2, \dots, x_n)$ can again be represented as an infinite number of linear inequalities $y \geq \langle a, x \rangle - \phi(a)$, $a = (a_1, \dots, a_n) \in \mathbb{R}^n$, where

$\langle a, x \rangle$ is the *inner product* defined as $\langle a, x \rangle = a_1x_1 + a_2x_2 + \cdots + a_nx_n$, and the function $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$ is defined by

$$\phi(a) := \max_{x \in \mathbb{R}^n} (\langle a, x \rangle - f(x))$$

and is called the Legendre transform of f . This trick is extremely useful in convex optimization.

The Characterization of the Legendre Transform

The theorem below, proved in [20], shows that the Legendre transform is, up to linear terms, the *only* transformation which has the useful properties (P1) and (P2). To formulate it, we need a few more definitions. A set $S \subset \mathbb{R}^n$ is called *closed* if $x_n \in S$, $\forall n$ and $\lim_{n \rightarrow \infty} x_n = x$ implies that $x \in S$. For example, $[0, 1]$ is a closed set, while $(0, 1]$ is not, because it contains a sequence $x_n = 1/n$, $n = 1, 2, \dots$, but not its limit point 0. A function $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$ is called *lower-semicontinuous* if its epigraph is a closed set in \mathbb{R}^{n+1} . Let $\mathcal{C}(\mathbb{R}^n)$ be the set of all lower-semi-continuous convex functions $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$. A transformation $B : \mathbb{R}^n \rightarrow \mathbb{R}^n$, sending (x_1, x_2, \dots, x_n) to (y_1, y_2, \dots, y_n) , is called *linear* if $y_j = b_{1j}x_1 + b_{2j}x_2 + \cdots + b_{nj}x_n$, $j = 1, 2, \dots, n$, for some real coefficients b_{ij} , $i = 1, 2, \dots, n$, $j = 1, 2, \dots, n$, *symmetric* if $b_{ij} = b_{ji}$, $\forall i, j$, and *invertible* if $B(x) \neq 0$ whenever $x \neq 0$.

Theorem 9.5 Assume that a transform $T : \mathcal{C}(\mathbb{R}^n) \rightarrow \mathcal{C}(\mathbb{R}^n)$ (defined on the whole domain $\mathcal{C}(\mathbb{R}^n)$) satisfies (P1) and (P2), that is, $TTf = f$ and $f \leq g$ implies $Tf \geq Tg$. Then T is essentially the Legendre transform L . Namely, there exists a constant $C_0 \in \mathbb{R}$, a vector $v_0 \in \mathbb{R}^n$, and an invertible symmetric linear transformation B such that

$$(Tf)(x) = (Lf)(Bx + v_0) + \langle x, v_0 \rangle + C_0, \quad \forall x \in \mathbb{R}^n.$$

Reference

S. Artstein-Avidan and V. Milman, The concept of duality in convex analysis, and the characterization of the Legendre transform, *Annals of Mathematics* **169**-2, (2009), 661–674.

9.6 The Solution of the Ten Martini Problem

Operators and Operations Between Them

Familiar functions like $f(x) = x$, $f(x) = 2x$, or $f(x) = x^2$, map real numbers to real numbers. In geometry, we study motions of the plane, like rotations or reflections,

which can be viewed as functions which map points of the plane to other points. If each point is given by two real coordinates, such functions map pairs of real numbers into pairs. For example, the function $f(x, y) = (-x, -y)$ represents reflection with respect to the point $(0, 0)$.

Here, we consider functions that map infinite sequences of numbers to infinite sequences. Such functions are called *operators*. Every operator T takes an infinite sequence $x = (x_1, x_2, x_3, \dots)$ as an input, and transforms it into another infinite sequence $y = (y_1, y_2, y_3, \dots)$, which we will also denote by $T(x)$. The simplest operator, usually denoted by I , is the identity operator, which sends every sequence to itself, that is, $I(x) = x$ for all sequences x . A little less trivial is the multiplication by constant operator, which just multiplies every term of the sequence by the same constant λ , that is, transforms every infinite sequence $x = (x_1, x_2, x_3, \dots)$ into the sequence $\lambda x = (\lambda x_1, \lambda x_2, \lambda x_3, \dots)$. Another example is the shift operator, which we denote by S , which transforms every sequence $x = (x_1, x_2, x_3, \dots)$ into the sequence $S(x) = (0, x_1, x_2, x_3, \dots)$.

Operators can be added together and multiplied by constants. The product of any operator T and constant $\lambda \in \mathbb{R}$ is an operator, denoted λT , which maps every sequence x into the sequence $\lambda T(x)$. For example, λI is just the “multiplication by λ ” operator, while λS is the operator transforming every sequence (x_1, x_2, x_3, \dots) into the sequence $(0, \lambda x_1, \lambda x_2, \lambda x_3, \dots)$. The sum of two operators T_1 and T_2 , denoted $T_1 + T_2$, is the operator which maps every sequence x to the sequence $T_1(x) + T_2(x)$, where the sequences are added element-wise. The difference $T_1 - T_2$ is just $T_1 + (-1) \cdot T_2$. For example, the operator $\lambda I - S$ maps every sequence (x_1, x_2, x_3, \dots) to the sequence $(\lambda x_1, \lambda x_2 - x_1, \lambda x_3 - x_2, \dots)$.

The Norm of an Infinite Sequence

For any vector in the plane with coordinates (x, y) its length is $\sqrt{x^2 + y^2}$; for a vector (x, y, z) in three-dimensional space the length is $\sqrt{x^2 + y^2 + z^2}$. Can we define the “length” of an infinite sequence in a similar way? For some sequences, like $(1, 2, 3, \dots)$, this seems to be difficult, but for others, like $(1, 1/2, 1/4, 1/8, \dots)$, a similar formula works. If we square all the “coordinates” and add them together, we get the expression $1^2 + (1/2)^2 + (1/4)^2 + (1/8)^2 + \dots$, which is the same as $1 + 1/4 + (1/4)^2 + (1/4)^3 + \dots$. In general, for any q , the sum $1 + q + q^2 + \dots + q^n$ is equal to $\frac{1}{1-q} - \frac{q^n}{1-q}$. If $q \in (0, 1)$, the term $\frac{q^n}{1-q}$ becomes smaller and smaller, hence the whole sum becomes closer and closer to $\frac{1}{1-q}$. In this case, we say that the infinite sum $1 + q + q^2 + \dots$ converges to $\frac{1}{1-q}$ and write $1 + q + q^2 + \dots = \frac{1}{1-q}$. In our case, $q = 4$, and $1 + 1/4 + (1/4)^2 + (1/4)^3 + \dots = \frac{1}{1-1/4} = \frac{4}{3}$, hence the infinite sequence $(1, 1/2, 1/4, 1/8, \dots)$ has finite “length” $\sqrt{\frac{4}{3}}$. In general, the set of all sequences $x = (x_1, x_2, x_3, \dots)$ with finite sum $x_1^2 + x_2^2 + x_3^2 + \dots$ is denoted l^2 , and the square root of this sum is denoted $\|x\|$ and is called the *norm* of x .

For example, the sequence $(1, 1/2, 1/4, 1/8, \dots)$ belongs to the set l^2 , while the sequence $(1, 2, 3, \dots)$ does not.

Bounded Linear Operators on l^2

All the operators T considered above have the property that if the input x belongs to l^2 , then so does the output $T(x)$. For example, if the sum $x_1^2 + x_2^2 + x_3^2 + \dots$ converges to some finite number A , then the sum $(\lambda x_1)^2 + (\lambda x_2)^2 + (\lambda x_3)^2 + \dots$ converges to a finite number $\lambda^2 A$, hence $x \in l^2$ implies that $\lambda x \in l^2$. Also, if $x_1^2 + x_2^2 + x_3^2 + \dots$ converges to A , then $0^2 + x_1^2 + x_2^2 + x_3^2 + \dots$ converges to A as well. In other words, $x \in l^2$ implies that $S(x) \in l^2$. From now on, we consider only operators T such that $T(x) \in l^2$ whenever $x \in l^2$.

An operator T is called *linear* if $T(x + y) = T(x) + T(y)$ for all sequences $x, y \in l^2$. It is easy to check that operators I , λI , and S are linear. A linear operator is called *bounded* if there is a constant $M > 0$ such that $\|T(x)\| \leq M \|x\|$ for all sequences $x \in l^2$. For example, operators λI and S satisfy this property with $M = |\lambda|$ and $M = 1$, respectively.

The Spectrum of a Bounded Operator

An operator T is called *invertible* if for any sequence $y \in l^2$ there exists a unique sequence $x \in l^2$ such that $T(x) = y$. For example, the operator I is trivially invertible, with $x = y$. More generally, the operator λI is invertible for every $\lambda \neq 0$, with $x = (1/\lambda)y$. On the other hand, the operator S is not invertible, because for any sequence $y = (y_1, y_2, y_3, \dots)$ with $y_1 \neq 0$ there is no x with $S(x) = y$.

What about the operator $\lambda I - S$, mapping (x_1, x_2, x_3, \dots) to $(\lambda x_1, \lambda x_2 - x_1, \lambda x_3 - x_2, \dots)$? For $\lambda = 0$, it reduces to $-S$ and is not invertible. For $\lambda \neq 0$ and any $y = (y_1, y_2, y_3, \dots)$, the equation $T(x) = y$ implies that $\lambda x_1 = y_1$, $\lambda x_2 - x_1 = y_2$, $\lambda x_3 - x_2 = y_3$, and so on. From the first equation, $x_1 = y_1/\lambda$; from the second one, $x_2 = (y_2 + x_1)/\lambda = y_2/\lambda + y_1/\lambda^2$; from the third one, $x_3 = y_3/\lambda + y_2/\lambda^2 + y_1/\lambda^3$, and so on. Continuing in this way, we can restore $x = (x_1, x_2, x_3, \dots)$ uniquely, hence $\lambda I - S$ is invertible for any $\lambda \neq 0$.

In general, the set of all real numbers λ such that the operator $\lambda I - T$ is *not* invertible is called the *spectrum* of a bounded operator T . For example, we have just proved that the spectrum of the shift operator S consists of one number $\lambda = 0$. It is also easy to see that the spectrum of I is one number $\lambda = 1$. In general, however, the spectrum can have a much more complicated structure, and the study of the spectral properties of linear operators is an important area of mathematical research.

The Almost Mathieu Operator and Its Spectrum

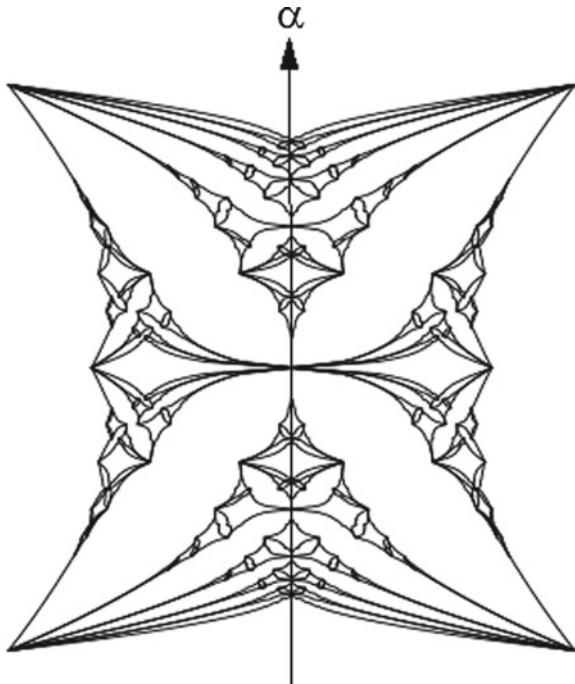
One important operator arising from applications in physics is the so-called *almost Mathieu operator*, which depends on three real parameters $\lambda \neq 0$, α , and θ , and transforms every sequence $x = (x_1, x_2, x_3, \dots)$ into the sequence $y = (y_1, y_2, y_3, \dots)$ according to the formula

$$y_n = x_{n+1} + x_{n-1} + 2\lambda \cos(2\pi(\theta + n\alpha))x_n.$$

The study of the spectrum of this operator, motivated by physical applications, has kept mathematicians busy for several decades. Experiments show that it has a complicated structure. For example, if one fixes λ and θ , and depicts the spectrum of the almost Mathieu operator for various α , one usually gets a fractal-like picture as in Fig. 9.6.

If α is a rational number (that is, $\alpha = p/q$ for some integers p, q), then the spectrum consists of the union of q intervals. For irrational α , it was conjectured that the spectrum is a so-called *Cantor set*. The most well-known example of a Cantor set is when you start with the interval $[0, 1]$, remove the middle third $(1/3, 2/3)$, then remove middle thirds $(1/9, 2/9)$ and $(7/9, 8/9)$ of the remaining intervals $[0, 1/3]$ and $[2/3, 1]$, then the middle thirds of the remaining four intervals, and so on, up to infinity. The set C of all points which survive is a Cantor set, see Sect. 1.4 for a

Fig. 9.6 Spectra of the almost Mathieu operator for various α



more detailed discussion. Of course, we have some flexibility in this construction, e.g. we can remove some other fixed proportion of every interval at every step, or even different proportions at every step, etc. However, all the sets constructed in this way are *homeomorphic*, that is, for every pair of them, there is a continuous invertible function which maps one set into the other. In general, a Cantor set is any set homeomorphic to the one described above.

The conjecture that, for every irrational α , the spectrum of the almost Mathieu operator is a Cantor set, was proposed by Azbel [26] in 1964. In 1981, Mark Kac offered ten martinis for anyone who could prove or disprove it, and since then the problem has been known as “the Ten Martini Problem”. Despite many partial results for some special irrational α , the general case was open until 2009, when the final positive resolution by Artur Avila and Svetlana Jitomirskaya appeared [24].

Theorem 9.6 *The spectrum of the almost Mathieu operator is a Cantor set for all irrational α and for all θ and all $\lambda \neq 0$.*

Reference

A. Avila and S. Jitomirskaya, The Ten Martini Problem, *Annals of Mathematics* **170**-1, (2009), 303–342.

9.7 A Linear Time Algorithm for Edge-Deletion Problems

The Party Organization Problem

In Sect. 5.11 we discussed the “party organization problem”: if some of your guests do not like each other, what is the minimal number of tables you need to be able to guarantee that no pair of enemies share a table? The correct mathematical language in which to study this problem is graph theory: we can represent the guests as points in the plane (vertices), and join any two vertices by a line (edge) if and only if the corresponding guests are enemies. We can also represent the tables as colours, and ask what is the minimal number of colours we need to colour the vertices of the graph in such a way that no two vertices connected by an edge have the same colour. A set of vertices, some of which are connected by edges, is called a *graph*, and the minimal number of colours in the problem described above is called the *chromatic number* of the graph. A graph with chromatic number at most k is called k -colourable.

In most restaurants, however, we have no control over the number of available tables. The restaurant may inform you that they have k big tables, where k is a small fixed number such as $k = 2$, and, in this case, it may be impossible to seat *every* pair of enemies at different tables. For example, if you have $k = 2$ tables and $n = 4$ guests Anna, Bob, Claire, and David, such that Anna and Bob dislike everyone else, including each other (but Claire and David are friends), you can start by putting enemies Anna and Bob at different tables, but then you need to put Claire either near

Anna or near Bob, creating an unhappy pair of enemies at the same table. Moreover, you then need to put David either near Anna or near Bob as well, creating another unhappy pair.

While a perfect solution in this case is impossible, you can still do better than the way described above. Namely, you can put Anna and Bob at table 1, and Claire and David at table 2. In this case you still have a pair of enemies (Anna and Bob) seating near the same table, but at least you have *one* such pair, not *two*!

Edge Removal and Graph Colouring

In general, our problem is to seat n guests at k tables such that the number of pairs of enemies at the same table is as small as possible. In the language of graph theory, we have a graph G with n vertices, and the problem is to colour the vertices in k colours such that the number of edges which join vertices of the same colour is minimal. The same question can be formulated slightly differently: what is the minimal number of edges we should *remove* from G to make it k -colourable? In the example above, we had a graph with vertices A , B , C , and D (A —Anna, B —Bob, C —Claire, D —David) and edges AB , AC , AD , BC , BD , see Fig. 9.7a. This graph is not 2-colourable, but, after removing just one edge AB , it becomes 2-colourable with vertices A , B coloured white, while C and D are coloured black. This colouring represents the way the guests should sit to create just one unhappy pair Anna-Bob—the pair whose edge was removed. In general, if the graph is k -colourable after removing m edges, then this colouring represents a guest distribution with exactly m unhappy pairs.

Removing Edges to “Kill” Triangles or Squares

Similar problems in the form “What is the minimal number of edges we should remove from a graph to make it (something)?” arise in many subareas of graph theory and its applications. For example, a “triangle” is a triple of vertices A , B , C , such that all of them are connected by edges (that is, the graph contains edges AB , BC , and CA). A graph G is called *triangle-free* if it contains no triangles. For any graph G with triangles, we may ask what is the minimal number of edges we should remove from G to make it triangle-free. If a graph G contains m triangles,

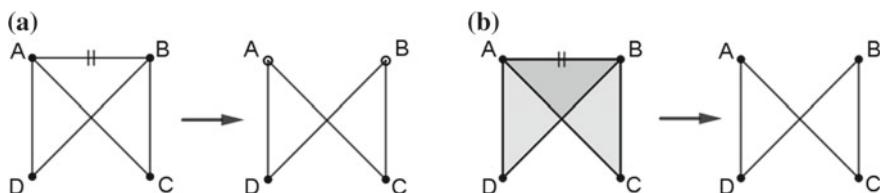


Fig. 9.7 Making the Anna-Bob-Claire-David graph 2-colourable and triangle-free

then removing m edges (one in each triangle) would surely work, but sometimes we can do better. For example, the Anna-Bob-Claire-David graph in Fig. 9.7b contains two triangles, ABC and ABD , but removing a single edge AB destroys them both, and makes it triangle-free after just one edge deletion.

In a similar way we may ask how many edges we should remove to make the graph G *square-free*, that is, containing no four vertices A, B, C, D such that AB, BC, CD and DA are edges in G . More generally, we may aim to avoid any fixed configuration like this.

The General Edge Removal Problem for Monotone Properties

In general, a property P of a graph is called *monotone* if it is preserved after removal of vertices and edges of G . For example, if a graph G is k colourable, then, after removing any vertex or any edge from it, it obviously remains k colourable, because the same colouring works. Similarly, if G is triangle-free, then after removing any vertex or edge from G it obviously remains triangle-free. Hence, the properties of being k colourable or being triangle-free are examples of monotone properties. The same is true for the property of being square-free, and for many other graph properties of theoretical and practical interest.

In this general setting, our problem is formulated as follows:

- (*) Given a monotone property P and arbitrary graph G , what is the minimal number of edge deletions needed to turn G into a graph satisfying P ?

Problem (*) is very difficult to solve exactly. A naïve approach would be to just try all possible edge deletions, but this works only for small graphs. In a graph with m edges, there are m ways to delete the first edge, $m - 1$ ways to delete the second one, and so on, so there are not much less than m^k ways to delete k edges. For a large graph with $m = 1000$ edges, and $k = 10$, $m^k = 10^{30}$. Even for a supercomputer performing 10^{16} operations per second, it would take more than three millions years to perform 10^{30} operations. Moreover, for some graph properties P it is not easy even to check if the initial graph G satisfies P . For example, this is the case if P is the k -colourability property for $k \geq 3$, because there are a huge number of possible colourings, and it could take ages to check if there is one that works.

An Approximate Solution

Because problem (*) is difficult to solve exactly, an important question is whether it is possible to efficiently find at least an *approximate* solution to it. This is what the following theorem of Noga Alon, Asaf Shapira, and Benny Sudakov [11] is about.

Theorem 9.7 *For any fixed $\varepsilon > 0$ and any monotone property P , there is a constant C (depending on ε and P) and an algorithm which, given a graph G with n vertices and m edges, finds an approximate solution to (*) to within an additive error εn^2 after performing at most $C(n + m)$ operations.*

In other words, if the exact optimal answer to $(*)$ is $k(P, G)$, the algorithm will return an answer $k'(P, G)$ such that $|k'(P, G) - k(P, G)| \leq \varepsilon n^2$. In many cases, this is a reasonable approximation. Indeed, the number of pairs of vertices of G is a bit less than $n^2/2$ (the exact formula is $n(n-1)/2$). If about half of all pairs are connected by edges, then there are about $m \approx n^2/4$ edges. If $\varepsilon = 0.001$ or so, then the error εn^2 is much less than the total number of edges. For example, if $n = 1000$ and $m = n^2/4 = 250,000$, then the solution to $(*)$ may be anything between 0 and 250,000, while the algorithm outputs the number k' , and guarantees that the answer is between $k' - 1000$ and $k' + 1000$.

Can we develop an algorithm with an even better approximation guarantee, e.g. with additive error proportional to $n^{1.99}$ instead of n^2 , or at least to $n^{2-\delta}$ for some $\delta > 0$? The authors prove that this is possible if there is a 2-colourable graph that does *not* satisfy P. For example, this is the case for the property of being square-free, because the square itself (a graph with four vertices A, B, C, D such that AB, BC, CD and DA are edges) is clearly not square-free but is 2-colourable (to see this, colour A and C green and B and D blue).

On the other hand, if P is a property such that all 2-colourable graphs satisfy P (this is the case if P is the property of being k -colourable for $k \geq 2$, or triangle-free), then the authors provide very strong evidence that, for any $\delta > 0$, no efficient algorithm with approximation guarantee $n^{2-\delta}$ exists.

Reference

N. Alon, A. Shapira, and B. Sudakov, Additive approximation for edge-deletion problems, *Annals of Mathematics* **170**-1, (2009), 371–411.

9.8 A Characterization of Stability-Preserving Linear Operators

Polynomials and Their Roots

A (real) polynomial is any function of the form

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0,$$

where a_0, a_1, \dots, a_n are some real coefficients. If $a_n \neq 0$, we say that the polynomial $P(x)$ has degree n . For example, polynomials of degree 0 are just constant functions $P(x) = a_0$, $a_0 \neq 0$, polynomials of degree 1 are linear functions $P(x) = a_1 x + a_0$, $a_1 \neq 0$, polynomials of degree 2 are quadratic functions, $P(x) = a_2 x^2 + a_1 x + a_0$, $a_2 \neq 0$, and so on.

A (real) root of a polynomial $P(x)$ is a real solution to the equation $P(x) = 0$. If the polynomial $P(x)$ can be written as $P(x) = (x - a)Q(x)$, where $Q(x)$ is another real polynomial, then $P(a) = (a - a)Q(a) = 0$, hence a is a root of $P(x)$. If, moreover, $P(x)$ can be written as $P(x) = (x - a)^k Q(x)$, then we say that a is a root of $P(x)$

of multiplicity k , and then the convention is that the root a should be counted k times while counting the roots of $P(x)$. For example, we say that the polynomial $P(x) = (x - 1)(x - 3)^2$ has *three* roots: 1, 3, and 3 again.

Stable Polynomials

Polynomials of degree 0, 1, and 2 have at most 0, 1, and 2 real roots, respectively. This is not a coincidence. There is an (easy) mathematical theorem stating that *any* polynomial of degree n has at most n real roots. Polynomials of degree n which have n real roots (that is, the maximal possible number of roots), are called *stable*, or *hyperbolic*. By convention, we consider the special polynomial $P(x) = 0$ to be stable as well. For example, the polynomial $P(x) = 2x - 3$ is stable, because its degree is $n = 1$, and it has one root $x = 3/2$. The polynomial $P(x) = x^2 - 3x + 2$ has degree $n = 2$ and two roots $x = 1$ and $x = 2$, hence it is stable. The polynomial $P(x) = x^2 - 2x + 1$ has degree $n = 2$ and root $x = 1$ of multiplicity 2, which is counted as two roots, hence it is stable as well. However, the polynomial $P(x) = x^2 + 1$ has degree $n = 2$ but no real roots at all, hence it is *not* stable.

Stability Under Differentiation

The *derivative* of a polynomial $P(x)$ is a polynomial, denoted $P'(x)$, which can be characterized using the following rules

- (i) $(P + Q)'(x) = P'(x) + Q'(x)$ for all polynomials P, Q ,
- (ii) $(aP)'(x) = aP'(x)$ for every polynomial P and constant $a \in \mathbb{R}$,
- (iii) $(x^k)' = kx^{k-1}$ for all $k \geq 0$.

For example, let us calculate the derivative of the polynomial $P(x) = x^2 - 3x + 2$. Rules (i) and (ii) imply that $P'(x) = (x^2 - 3x + 2)' = (x^2)' + (-3x)' + (2)' = (x^2)' - 3(x)' + 2(1)'$. By (iii), $(x^2)' = 2x$, $(x)' = 1$, and $(1)' = (x^0)' = 0$, hence $P'(x) = 2x - 3 \cdot 1 + 2 \cdot 0 = 2x - 3$. One easy but useful theorem is that the derivative of any stable polynomial is stable. In other words, we say that *stability is preserved under differentiation*. For example, $P(x) = x^2 - 3x + 2$ is a stable polynomial (degree 2, and 2 roots), and its derivative $P'(x) = 2x - 3$ is stable as well (degree 1, and 1 root).

Stability After Multiplication

Any individual polynomial, say $P(x) = x^2 - 3x + 2$, is a function transforming real numbers into real numbers, e.g. the number $x = 4$ is transformed into $P(4) = 4^2 - 3 \cdot 4 + 2 = 6$. In contrast, differentiation is an example of an “operation” transforming polynomials into polynomials, e.g. $x^2 - 3x + 2$ is transformed into $2x - 3$. Another example of such an “operation” is multiplication by any fixed polynomial

$Q(x)$, say, $Q(x) = x - 5$. In this case, the polynomial $P(x) = x^2 - 3x + 2$ is transformed into $(x^2 - 3x + 2)(x - 5) = x^3 - 8x^2 + 17x - 10$. The roots of the “transformed” polynomial are the same as the roots of the original one plus the roots of Q . In particular, this implies that $Q(x) \cdot P(x)$ is a stable polynomial whenever $P(x)$ and $Q(x)$ are stable. In other words, *stability is preserved after multiplication by a stable polynomial*.

Linear Operators Transforming Polynomials

In general, a *linear operator* is any “operation” T transforming polynomials into polynomials which satisfies properties (i) and (ii) above, that is, (i) $T(P + Q) = T(P) + T(Q)$ for all polynomials P, Q , and (ii) $T(aP) = aT(P)$ for every polynomial P and constant $a \in \mathbb{R}$. This implies that $T(x^2 - 3x + 2) = T(x^2) - 3T(x) + 2T(1)$, and, more generally,

$$T(a_n x^n + \dots + a_1 x + a_0) = a_n T(x^n) + \dots + a_1 T(x) + a_0 T(1),$$

that is, to define T , it suffices to define $T(x^k)$ for all $k \geq 0$. For example, if $T(x^k) = kx^{k-1}$, $k \geq 0$, then T is differentiation, while $T(x^k) = x^k \cdot Q(x)$, $k \geq 0$, implies that T is just multiplication by a fixed polynomial $Q(x)$. In general, however, T can be arbitrarily complicated: for example, it may be that $T(x^k) = x^3 - 4x$ for even k while $T(x^k) = x^2 - 1$ for odd k . In this case, $T(x^2 - 3x + 2) = (x^3 - 4x) - 3(x^2 - 1) + 2(x^3 - 4x) = 3x^3 - 3x^2 - 12x + 3$, and, in general, for any polynomial P , $T(P)$ has the form $a(x^3 - 4x) + b(x^2 - 1)$ for some constants a, b . It is not difficult to verify that the polynomial $a(x^3 - 4x) + b(x^2 - 1)$ is stable for all a, b , hence, in this case, $T(P)$ is stable for all P .

Which Linear Operators Preserve Stability?

One of the long-standing fundamental problems in the theory of stable polynomials was to “characterize” *all* linear operators T which preserve stability, that is, such that $T(P)$ is always a stable polynomial whenever P is stable. Here, by “characterization” we mean simple-to-check necessary and sufficient conditions. In 1914, such conditions were derived by Pólya and Schur [308] for operators of the form $T(x^k) = \lambda^k x^k$, $k \geq 0$, where $\lambda_0, \lambda_1, \lambda_2, \dots$ is a given sequence of numbers. Since then, there have been many similar results covering very special transformations T , but almost no progress for general T , until the question was fully resolved [67] in 2009!

To formulate the result, we need some more definitions. We say that stable polynomials $P(x)$ and $Q(x)$ are *interlacing* if either $\alpha_1 \leq \beta_1 \leq \alpha_2 \leq \beta_2 \leq \dots$ or $\beta_1 \leq \alpha_1 \leq \beta_2 \leq \alpha_2 \leq \dots$, where $\alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_n$ and $\beta_1 \leq \beta_2 \leq \dots \leq \beta_m$ are roots of $P(x)$ and $Q(x)$, respectively. Note that this condition may be satisfied only if n and m differ by at most one. For example, polynomials $P(x) = x^3 - 4x$ and

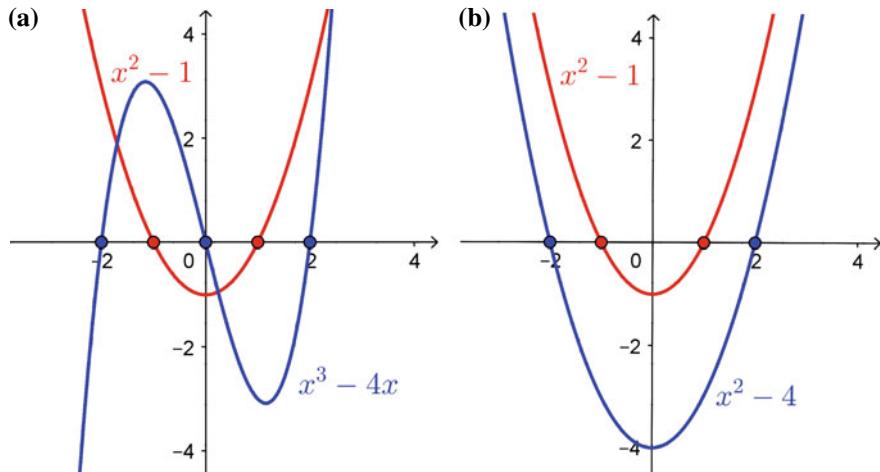


Fig. 9.8 Interlacing and non-interlacing polynomials

$Q(x) = x^2 - 1$ are interlacing, because their roots are $-2, 0, 2$ and $-1, 1$, respectively. In contrast, polynomials $x^2 - 4$ and $x^2 - 1$ are not interlacing, see Fig. 9.8. Stable interlacing polynomials R and Q have the property that $aR(x) + bQ(x)$ is stable for all a, b . In particular, if T is a linear operator such that $T(P)$ has the form $aR(x) + bQ(x)$ for all P , then $T(P)$ is stable for all P .

A polynomial $P(x, y)$ in two variables x, y is the sum of any finite number of terms of the form $ax^k y^m$, where $a \in \mathbb{R}$ and k, m are non-negative integers. $P(x, y)$ is called *stable* if $Q(t) = P(a + bt, c + dt)$ is a stable polynomial in one variable t for any real a, b, c, d such that $b > 0$ and $d > 0$. For example, $P(x, y) = x + y$ is stable because in this case $Q(t) = (a + c) + (b + d)t$ has degree 1 and one root $t = -(a + c)/(b + d)$.

For every linear operator T , let S_T be the linear operator transforming polynomials in two variables into other polynomials in two variables according to the rule $S_T(x^k y^l) = T(x^k)y^l$. For example, if T is differentiation, then S_T is known as the partial derivative with respect to x , and is calculated by the rule $S_T(x^k y^l) = kx^{k-1}y^l$, for example, $S_T(x^3 y + x^2 y^2) = 3x^2 y + 2x y^2$.

Theorem 9.8 *A linear operator T , transforming polynomials into polynomials, preserves stability if and only if*

- (a) $T(x^k) = a_k P(x) + b_k Q(x)$, where $a_k, b_k, k = 0, 1, 2, \dots$ are real numbers, and $P(x)$ and $Q(x)$ are some fixed (independent of k) stable interlacing polynomials; or
- (b) $S_T[(x + y)^k]$ is a stable polynomial (in 2 variables) for all $k = 0, 1, 2, \dots$; or
- (c) $S_T[(x - y)^k]$ is a stable polynomial (in 2 variables) for all $k = 0, 1, 2, \dots$.

Reference

J. Borcea and P. Brăndén, Pólya–Schur master theorems for circular domains and their boundaries, *Annals of Mathematics* **170**-1, (2009), 465–492.

9.9 On the Gaps Between Primes

The Average Distance Between Consecutive Primes

Primes are natural numbers with exactly two divisors, like 2, 3, 5, 7, 11, 13, 17, 19, 23, Because all even numbers n greater than 2 have at least three divisors (1, 2, and n), 2 is the only even prime number. This implies that the pair 2, 3 is the only pair of consecutive prime numbers.

The pairs $p = 3, q = 5$, or $p = 5, q = 7$, or $p = 11, q = 13$, and so on, are examples of pairs of primes p and q such that $q - p = 2$. The famous twin primes conjecture states that there are infinitely many such pairs, and it is one of the oldest unsolved problems in mathematics. A “naïve” reason why this conjecture may be hard to prove is that, if we study the sequence of primes further and further, the *average* distance between consecutive primes becomes larger and larger. The famous prime number theorem states that, for any large N , there are about $\frac{N}{\ln N}$ primes less than N . Hence, the average distance between consecutive primes is approximately $\ln N$. For $N = 10^{100}$, this implies that the average distance between 100-digit primes is about 230. Of course, this does not mean that this distance is exactly 230 in all cases: for some pairs of consecutive primes it is larger, while for some pairs it is smaller.

Pairs of Primes at Distance Much Lower Than the Average

Let p_n denote the n -th prime, so that $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, $p_4 = 7$, and so on. The prime number theorem states that $p_{n+1} - p_n$ is, on average, about $\ln p_n$. The twin primes conjecture states that $p_{n+1} - p_n = 2$ for infinitely many values of n . To make progress towards it, can we at least prove that $p_{n+1} - p_n$ is less than average infinitely often? That is, given some $\varepsilon \in (0, 1)$, can we prove that

$$(*) \quad p_{n+1} - p_n \leq \varepsilon \ln p_n \text{ for infinitely many values of } n?$$

In 1926, Hardy and Littlewood proved $(*)$ for $\varepsilon = \frac{2}{3}$, assuming an unproven conjecture called the Generalized Riemann Hypothesis. Unconditionally, Erdős [140] proved in 1940 that $(*)$ holds for *some* $\varepsilon \in (0, 1)$, but he did not provide an explicit value. In 1954, Ricci [320] proved $(*)$ for $\varepsilon = \frac{15}{16}$, and then there was a long chain of improvements, with the best result before 2009 being a 1988 theorem of Maier, [260] proving that $(*)$ holds for $\varepsilon \approx 0.2484$.

In 2009, Goldston, Pintz, and Yıldırım [165] proved the following theorem.

Theorem 9.9 *For any $\varepsilon > 0$, there exists infinitely many values of n such that*

$$p_{n+1} - p_n \leq \varepsilon \ln p_n.$$

Theorem 9.9 states that (*) holds for *any* $\varepsilon > 0$, no matter how small. In the authors' words, "there exist consecutive primes which are closer than any arbitrarily small multiple of the average spacing".

While Theorem 9.9 is huge progress compared to the previous results, it is still far from confirming the twin primes conjecture.

Primes in Arithmetic Progressions: Dirichlet's Theorem

In addition to proving Theorem 9.9, Goldston, Pintz, and Yildirim provided an excellent idea for further progress, which is based on the distribution of primes in arithmetic progressions. An *arithmetic progression* with first term a and difference q is a sequence of the form

$$a, a + q, a + 2q, a + 3q, \dots, a + kq, \dots \quad (9.10)$$

For example, $4, 10, 16, 22, 28, \dots$ is an arithmetic progression with $a = 4$ and $q = 6$. This particular arithmetic progression contains no primes at all, because all terms in it are divisible by 2. In general, if there is a number $r > 1$ such that both a and q are divisible by r , then all terms in (9.10) are divisible by r , hence it contains no primes at all, or possibly one prime which is equal to r . If there are no such r , the numbers a and q are called *relatively prime*. For example, 4 and 6 are not relatively prime, because they are both divisible by $r = 2$, while $a = 3$ and $q = 4$ are relatively prime. Dirichlet's famous theorem states that an arithmetic progression (9.10) contains infinitely many primes whenever a and q are relatively prime. For example, with $a = 3$ and $q = 4$, this implies that the sequence $S_1 = 3, 7, 11, 15, 19, 23, 27, 31, 35, \dots$ contains infinitely many primes, while with $a = 1$ and $q = 4$, we conclude that the sequence $S_2 = 1, 5, 9, 13, 17, 21, 25, 29, 33, \dots$ contains infinitely many primes as well.

Primes in Arithmetic Progressions: The Elliott–Halberstam Conjecture

In fact, all primes except for 2 belong either to S_1 or to S_2 , see Fig. 9.9, and we would expect that about half of them belong to each one. By the prime num-

1	5	9	13	17	21	25	29	33	37	41	45	49	53	57	61	65	69	73	77	81	85	89	93	...		
2	6	10	14	18	22	26	30	34	38	42	46	50	54	58	62	66	70	74	78	82	86	90	94	...		
3	7	11	15	19	23	27	31	35	39	43	47	51	55	59	63	67	71	75	79	83	87	91	95	...		
4	8	12	16	20	24	28	32	36	40	44	48	52	56	60	64	68	72	76	80	84	88	92	96	...		

Fig. 9.9 Primes of the form $4k + 1$ and $4k + 3$

ber theorem, for any large N , there are about $\frac{N}{\ln N}$ primes less than N , and we expect that about $\frac{N}{2 \ln N}$ of them belong to S_1 , and another $\frac{N}{2 \ln N}$ of them to S_2 . Also, by the prime number theorem, the product $\Pi(N)$ of all primes less than N (for example, $\Pi(12) = 2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 = 2310$) is approximately equal to e^N , where $e \approx 2.71828\dots$ is the base of the natural logarithm, and we expect that primes from S_1 and S_2 contribute approximately equally to this product, that is, $p_1 p_2 \dots p_k \approx p'_1 p'_2 \dots p'_m \approx \sqrt{e^N}$, where p_1, p_2, \dots, p_k and p'_1, p'_2, \dots, p'_m are primes less than N from S_1 and S_2 , respectively. Equivalently, $\ln(p_1 p_2 \dots p_k) \approx \ln(p'_1 p'_2 \dots p'_m) \approx \ln(\sqrt{e^N}) = N/2$. Or $g(N, 4, 3) \approx g(N, 4, 1) \approx N/2$, where $g(N, q, a)$ is the logarithm of the product of all primes less than N in the arithmetic progression (9.10). Similarly, for $q = 12$, we expect that primes are approximately uniformly distributed across four arithmetic progressions (9.10) with $a = 1, 5, 7, 11$ (these are all values of a less than 12 which are relatively prime with 12), and $g(N, 12, 1) \approx g(N, 12, 5) \approx g(N, 12, 7) \approx g(N, 12, 11) \approx N/4$. For general q , we expect that

$$g(N, q, a_1) \approx g(N, q, a_2) \approx \dots \approx g(N, q, a_{\phi(q)}) \approx \frac{N}{\phi(q)}, \quad (9.11)$$

where $a_1, a_2, \dots, a_{\phi(q)}$ are all values of a less than q which are relatively prime with q , and $\phi(q)$ denotes the number of such a , for example, $\phi(12) = 4$. Equation (9.11) is equivalent to saying that $|g(N, q, a_k) - \frac{N}{\phi(q)}|$ is “small” for $k = 1, 2, \dots, \phi(q)$, or, equivalently, that the function

$$h(N, q) := \max_{1 \leq k \leq \phi(q)} \left| g(N, q, a_k) - \frac{N}{\phi(q)} \right|$$

is “small”. To guarantee that this happens for *all* q not exceeding some value Q , we need to have a good upper bound for the function $H(N, Q) := \sum_{q \leq Q} h(N, q)$.

Elliott and Halberstam [127] conjectured that for any $v \leq 1$, any $A > 0$ and any $\varepsilon > 0$ there is a constant C such that

$$H(N, N^{v-\varepsilon}) \leq C \frac{N}{(\ln N)^A}, \quad \forall N. \quad (9.12)$$

Because $H(N, Q)$ is an increasing function in Q , (9.12) becomes harder to prove as v increases. The best theorem in this direction is the famous Bombieri–Vinogradov theorem proving (9.12) for $v \leq 0.5$.

Bounded Gaps Between Primes

Goldston, Pintz, and Yildirim proved that if (9.12) holds for *any* $v > 0.5$, even for $v = 0.50000001$, then there exist infinitely many values of n such that

$$p_{n+1} - p_n \leq B \quad (9.13)$$

where B is a constant depending only on v . In particular, if one can prove (9.12) for $v = 0.971$, then one can choose $B = 16$. This result already looks close to the twin primes conjecture, stating that the same statement holds with $B = 2$. However, (9.12) is currently known to hold only for $v \leq 0.5$, which is just a little bit less than needed!

In a later work, Zhang [408] observed that in fact an even weaker version of (9.12) implies (9.13), and was able to prove this weaker version, establishing (9.13) with $B = 70,000,000$. This was later improved by Maynard and others, and (9.13) is now known to hold with $B = 246$, see [309].

Reference

D. Goldston, J. Pintz, and C. Yıldırım, Primes in tuples I, *Annals of Mathematics* **170**-2, (2009), 819–862.

9.10 A Proof of the B. and M. Shapiro Conjecture in Real Algebraic Geometry

Bases and Linear Independence in \mathbb{R}^2

Any point A in the coordinate plane \mathbb{R}^2 can be described by two coordinates, x_A and y_A . Any two points B and A define a *vector* \mathbf{BA} , which is, geometrically, just an arrow connecting B with A . Algebraically, we say that the vector \mathbf{BA} has coordinates $(x_A - x_B, y_A - y_B)$, where (x_B, y_B) and (x_A, y_A) are coordinates of B and A , respectively. In particular, if $O = (0, 0)$ is the center of the coordinate plane, then the vector \mathbf{OA} has the same coordinates as A .

Vectors may be multiplied by constants using the rule $\alpha(x, y) = (\alpha x, \alpha y)$. If $A \neq O$, the set of all points M such that $\mathbf{OM} = \alpha\mathbf{OA}$, $\alpha \in \mathbb{R}$, is just a line passing through points O and A . If B is any point not on this line, then *any* vector \mathbf{OM} in the plane can be uniquely represented as a linear combination $\alpha\mathbf{OA} + \beta\mathbf{OB}$ of \mathbf{OA} and \mathbf{OB} , where addition is coordinate-wise. In this case, we say that the vectors \mathbf{OA} and \mathbf{OB} form a *basis* of the coordinate plane. For example, if A and B have coordinates $(2, 0)$ and $(1, 2)$, respectively, then any vector \mathbf{OM} with coordinates (x, y) can be uniquely represented as $(x, y) = \alpha(2, 0) + \beta(1, 2) = (2\alpha + \beta, 2\beta)$, see Fig. 9.10a, and the coefficients α and β in this representation are given by $\alpha = x/2 - y/4$ and $\beta = y/2$.

The condition “ B is not on the line OA ” is equivalent to “ $\mathbf{OB} \neq \alpha\mathbf{OA}$ for any $\alpha \in \mathbb{R}$ ”, or, equivalently, to $x_Ay_B - x_By_A \neq 0$. For example, for $(x_A, y_A) = (2, 0)$ and $(x_B, y_B) = (1, 2)$ this reduces to $2 \cdot 2 - 1 \cdot 0 \neq 0$. In this case, vectors with coordinates (x_A, y_A) and (x_B, y_B) are called *linearly independent*. In fact, two vectors in the plane form a basis if and only if they are linearly independent.

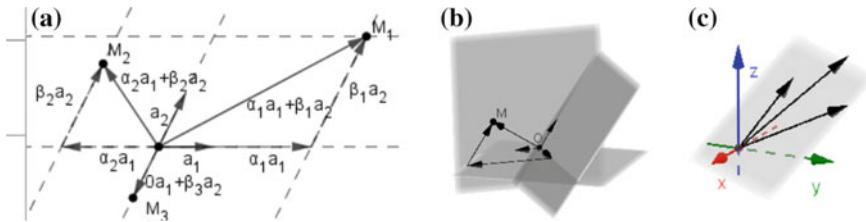


Fig. 9.10 Bases and linear independence in \mathbb{R}^2 and \mathbb{R}^3

Bases and Linear Independence in \mathbb{R}^3

Similarly, a vector in 3-dimensional space \mathbb{R}^3 is described by three coordinates (x, y, z) . More generally, a (real) n -dimensional vector is just a set of n real coordinates (x_1, x_2, \dots, x_n) . k such vectors $\mathbf{a}_1, \dots, \mathbf{a}_k$ are called *linearly independent* if there are no real numbers $\lambda_1, \dots, \lambda_k$, not all 0, such that $\lambda_1\mathbf{a}_1 + \dots + \lambda_k\mathbf{a}_k = 0$. For example, vectors $\mathbf{a}_1 = (2, 1, 0)$, $\mathbf{a}_2 = (-1, 2, 0)$ and $\mathbf{a}_3 = (-1, -1, 2)$ are linearly independent, and form a basis of \mathbb{R}^3 , see Fig. 9.10b, while vectors $\mathbf{a}_1 = (0, 4, 2)$, $\mathbf{a}_2 = (-2, 1, 2)$ and $\mathbf{a}_3 = (-2, 3, 3)$ are *not* linearly independent, because $0.5\mathbf{a}_1 + \mathbf{a}_2 - \mathbf{a}_3 = 0$. In fact, all linear combinations of these vectors form a plane, see Fig. 9.10c, and these vectors do *not* form a basis of \mathbb{R}^3 .

Polynomials in Real and Complex Variables

The notion of linearly independence can be studied not only for vectors, but for any mathematical “objects” which can be added and multiplied by constants, for example, polynomials. A real polynomial is any function of the form $P(x) = a_nx^n + a_{n-1}x^{n-1} + \dots + a_2x^2 + a_1x + a_0$, where a_0, a_1, \dots, a_n are some real coefficients. A *root* of a polynomial is a solution to the equation $P(x) = 0$. For example, if $n = 2$, $a_2 \neq 0$, the equation $P(x) = 0$ is a quadratic equation $a_2x^2 + a_1x + a_0 = 0$, whose solutions are given by the formula $x_{1,2} = \frac{-a_1 \pm \sqrt{a_1^2 - 4a_0a_2}}{2a_2}$. In particular, real solution(s) exist if and only if $a_1^2 - 4a_0a_2 \geq 0$. If $a_1^2 - 4a_0a_2 = -D$ for some $D > 0$, then $x_{1,2} = \frac{-a_1 \pm \sqrt{D}\sqrt{-1}}{2a_2} = \frac{-a_1 \pm \sqrt{D}i}{2a_2}$, where i is just a notation for the square root of -1 (which is not a real number). Numbers of the form $z = a + bi$ for some real a, b are called *complex* numbers, see e.g. Sect. 1.7 for details. The set of all complex numbers is usually denoted by \mathbb{C} .

As we have seen above, any quadratic equation with real coefficients *always* has complex roots. The fundamental theorem of algebra states that this remains correct for any equation of the form $P(z) = 0$, where $P(z)$ is a complex polynomial, that is, an expression of the form

$$P(z) = a_nz^n + a_{n-1}z^{n-1} + \dots + a_2z^2 + a_1z + a_0,$$

where z is a complex variable and a_0, a_1, \dots, a_n are complex coefficients.

Linear Independence and Bases for Polynomials

Two polynomials $P(z)$ and $Q(z)$ are called *linearly independent* if $P(z) \neq \alpha Q(z)$ and $Q(z) \neq \alpha P(z)$ for any complex number α . For example, $P(z) = iz^2 + (1 - i)$ and $Q(z) = -z^2 + (1 + i)$ are *not* linearly independent because $Q(z) = iP(z)$. In contrast, $P(z) = iz^2 + z$ and $Q(z) = z^2 + iz$ are linearly independent, because $\alpha(i z^2 + z) = z^2 + iz$, or $(\alpha i - 1)z^2 + (\alpha - i)z = 0$ implies that $\alpha i - 1 = 0$ and $\alpha - i = 0$, hence $\alpha = -i$ and $\alpha = i$, a contradiction.

More generally, k polynomials $P_1(z), P_2(z), \dots, P_k(z)$ are called *linearly independent* if there are no complex numbers $\lambda_1, \dots, \lambda_k$, not all 0, such that $\lambda_1 P_1(z) + \dots + \lambda_k P_k(z) = 0$. Let S be the set of polynomials which can be written as a linear combination of $P_1(z), P_2(z), \dots, P_k(z)$, that is,

$$S = \{P(z) \mid P(z) = \lambda_1 P_1(z) + \dots + \lambda_k P_k(z), \lambda_i \in \mathbb{C}, i = 1, \dots, k\}. \quad (9.14)$$

If $Q_1(z), Q_2(z), \dots, Q_k(z)$ are any other k linearly independent polynomials belonging to S , then the set S can be equivalently written as

$$S = \{P(z) \mid P(z) = \lambda_1 Q_1(z) + \dots + \lambda_k Q_k(z), \lambda_i \in \mathbb{C}, i = 1, \dots, k\}. \quad (9.15)$$

Any such set $Q_1(z), Q_2(z), \dots, Q_k(z)$ is called a *basis* for S .

Looking for a Simpler Basis

The representation (9.15) can sometimes be much simpler than the original representation (9.14). For example, if $k = 2$, $P_1(z) = (1 + i)z^2 + (1 - i)z + 2$, $P_2(z) = (1 - i)z^2 + (1 + i)z + 2$, then the set S in (9.14) consists of polynomials of the form

$$\begin{aligned} P(z) &= \lambda_1 P_1(z) + \lambda_2 P_2(z) \\ &= [\lambda_1(1 + i) + \lambda_2(1 - i)]z^2 + [\lambda_1(1 - i) + \lambda_2(1 + i)]z + [2\lambda_1 + 2\lambda_2]. \end{aligned}$$

Defining $\lambda'_1 := \lambda_1(1 + i) + \lambda_2(1 - i)$ and $\lambda'_2 := \lambda_1(1 - i) + \lambda_2(1 + i)$, we see that $2\lambda_1 + 2\lambda_2 = \lambda'_1 + \lambda'_2$, and $P(z) = \lambda'_1 z^2 + \lambda'_2 z + \lambda'_1 + \lambda'_2$, hence $S = \{P(z) \mid P(z) = \lambda'_1(z^2 + 1) + \lambda'_2(z + 1), \lambda'_1, \lambda'_2 \in \mathbb{C}\}$. Such a simplified representation is possible because S has a simple basis: $Q_1(z) = z^2 + 1$ and $Q_2(z) = z + 1$.

In general, it is convenient to represent S in (9.14) using as simple a basis as possible. In particular, what are sufficient conditions which guarantee the existence of a basis $Q_1(z), Q_2(z), \dots, Q_k(z)$ such that all polynomials $Q_i(z)$, $i = 1, \dots, k$ have only real coefficients?

Sufficient Conditions for the Existence of a Basis with Real Coefficients

To formulate the answer to this question, established in [285], we need more definitions. For any polynomial $P(z)$, its *derivative* is a polynomial $P'(z)$, uniquely determined by the rules (a) $(P + Q)'(z) = P'(z) + Q'(z)$ for all polynomials P, Q , (b) $(aP)'(z) = aP'(z)$ for every constant $a \in \mathbb{C}$, and (c) $(z^k)' = kz^{k-1}$ for $k = 0, 1, 2, \dots$. The second derivative of P , denoted $P^{(2)}(z)$, is the derivative of $P'(z)$, and so on. For example, for $P(z) = z^3 + iz^2 + 2z - i$, $P'(z) = 3z^2 + 2iz + 2$, $P^{(2)}(z) = 6z + 2i$, $P^{(3)}(z) = 6$, and $P^{(i)}(z) = 0$ for all $i \geq 4$.

For an arbitrary set of k polynomials $P_1(z), P_2(z), \dots, P_k(z)$, a complex number z^* is called a root of its *Wronskian* if the vectors $\mathbf{a}_1 = (P_1(z^*), P_2(z^*), \dots, P_k(z^*))$, $\mathbf{a}_2 = (P'_1(z^*), P'_2(z^*), \dots, P'_k(z^*))$, \dots , $\mathbf{a}_k = (P_1^{(k-1)}(z^*), P_2^{(k-1)}(z^*), \dots, P_k^{(k-1)}(z^*))$ are linearly dependent, that is, $\lambda_1 \mathbf{a}_1 + \dots + \lambda_k \mathbf{a}_k = 0$ for some complex numbers $\lambda_1, \dots, \lambda_k$, not all 0.

Theorem 9.10 *If all roots of the Wronskian of a set of polynomials $P_1(z), P_2(z), \dots, P_k(z)$ are real, then the set S defined in (9.14) has a basis consisting of polynomials with real coefficients.*

In the example above with $k = 2$, $P_1(z) = (1+i)z^2 + (1-i)z + 2$, $P_2(z) = (1-i)z^2 + (1+i)z + 2$, $z^* \in \mathbb{C}$ is a root of the Wronskian if vectors $\mathbf{a}_1 = (P_1(z^*), P_2(z^*))$ and $\mathbf{a}_2 = (P'_1(z^*), P'_2(z^*))$ are linearly dependent, which is the case if $P_1(z^*)P'_2(z^*) - P_2(z^*)P'_1(z^*) = 0$, where $P'_1(z^*) = 2(1+i)z^* + (1-i)$ and $P'_2(z^*) = 2(1-i)z^* + (1+i)$. This simplifies to $-4i(z^*)^2 - 8iz^* + 4i = 0$, or $(z^*)^2 + 2z^* - 1 = 0$. Because this equation has only real roots, Theorem 9.14 guarantees that S in (9.14) has a basis consisting of polynomials with real coefficients. As we have seen above, this is indeed the case, and the basis is $Q_1(z) = z^2 + 1$ and $Q_2(z) = z + 1$.

In fact, the $k = 2$ case of Theorem 9.14 was resolved in 2002, see Sect. 2.1, but the general case remained open until 2009. In its general form, Theorem 9.14 confirms a conjecture known as the “B. and M. Shapiro conjecture”, which has number of equivalent formulations, and many important consequences, especially in the field of mathematics called “real algebraic geometry”.

Reference

E. Mukhin, V. Tarasov, and A. Varchenko, The B. and M. Shapiro conjecture in real algebraic geometry and the Bethe ansatz, *Annals of Mathematics* **170**-2, (2009), 863–881.

9.11 Bounding Diagonal Ramsey Numbers

Looking for a Monochromatic Triangle

Assuming that there are six people in a room, can we always find either three people who all know each other or three people who all do not know each other? To analyse questions like this, it is convenient to represent people as points in the plane, and then connect the points by a blue line for any pair of people who know each other, and by a red line for any pair who do not know each other. Then we have 6 points, each pair connected by either a red or a blue line, and the question is can we always find either a red or a blue triangle?

Let us prove that the answer is “Yes, we always can”. From any point A we draw 5 lines, hence at least 3 of them should have the same colour, say, blue. Let A be connected by blue lines to points B, C , and D . If any of the lines BC, CD , or DB are blue, then we have a blue triangle (for example, if BC is blue, then the blue triangle is ABC , and so on). Otherwise all lines BC, CD , and DB are red, hence we have a red triangle BCD . In Fig. 9.11a you can see that you will get a monochromatic triangle after any colouring of BD .

What if we have just 5 people instead of 6? Then the answer to the same question is “No”. Let us label the people (and the corresponding points) A, B, C, D , and E , and let the lines AB, BC, CD, DE , and EA be blue, and the lines AC, CE, EB , BD and DA red, see Fig. 9.11b. It is easy to check that, in this case, neither a red nor a blue triangle exists. In fact, this colouring is “unique up to relabelling”, that is, in any set of 5 points connected by red or blue lines without red and blue triangles, we can always give the points names A, B, C, D , and E in such a way that the colouring becomes exactly as described above.

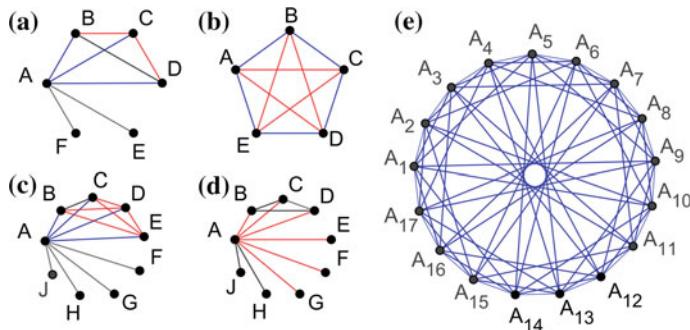


Fig. 9.11 Illustrations for **a** $R(3, 3) \leq 6$, **b** $R(3, 3) > 5$, **c** and **d** $R(4, 3) \leq 9$, **e** $R(4, 4) > 17$

Looking for a Red Triangle or Blue Quadruple

A slightly more difficult problem is to prove that, in a group of 9 people, we can always find either three who do not know each other, or *four* who know each other. In other words, if 9 points are connected by red or blue lines, then either there exists a blue triangle, or there are 4 points all connected by red lines, which we will call a *red quadruple*. Indeed, if every point is adjacent to exactly 3 blue lines, then the total number of blue lines is $9 \cdot 3/2$, which is not an integer, a contradiction. Hence, there is a point A adjacent either to at least 4 blue lines or to at most 2. In the first case, let A be connected by blue lines to points B, C, D and E , see Fig. 9.11c. If any pair of them (say, B and C) is connected by a blue line as well, then we have a blue triangle (in this case, ABC). Otherwise points B, C, D and E form a red quadruple. In the second case, A is connected by blue lines to at most 2 points, hence there are 6 points to which it is connected by red lines, see Fig. 9.11d. As we have proved above, out of these 6 points we can always select a triangle, call it BCD , which is either red or blue. If it is blue, we have found a blue triangle. If it is red, then $ABCD$ is a red quadruple.

Looking for a Monochromatic Quadruple

Similarly, we can prove that out of 18 points, connected by red or blue lines, we can always find either a red quadruple or a blue quadruple. Indeed, any point A is connected to 17 others, hence it is connected to at least 9 of them by lines of the same colour, say, blue. But we have just proved that in any set of 9 points we can always find either a blue triangle BCD (in which case $ABCD$ is a blue quadruple), or a red quadruple.

What if we have just 17 points, can we always find either a red or a blue quadruple? It turns out, we cannot. Let us label the points A_1, A_2, \dots, A_{17} , and position them in this order as the vertices of a regular 17-gon with unit side length. For any two points A, B , let $d(A, B)$ be the distance of the “shortest path” between them while travelling along the 17-gon: for example, $d(A_1, A_5) = 4$ with shortest path $A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow A_4 \rightarrow A_5$, while $d(A_2, A_{16}) = 3$ with shortest path $A_2 \rightarrow A_1 \rightarrow A_{17} \rightarrow A_{16}$. Let us colour the line AB blue if $d(A, B)$ is either 1, or 2, or 4, or 8, and red otherwise. In Fig. 9.11e only blue lines are depicted. Let us prove that there are no blue quadruples. Imagine we have one, with vertices (counter-clockwise) being A, B, C, D , and with $d(A, B) = a$, $d(B, C) = b$, $d(C, D) = c$, and $d(D, A) = d$. We can assume that $\max a, b, c, d = d$. Then either $a + b + c = d$ or $a + b + c + d = 17$. Because this is a blue quadruple, each of a, b, c, d are either 1, 2, 4, or 8, hence $a + b + c + d = 17$ is possible only if $d = 8$ and a, b, c are (in some order) 1, 4, 4. But then either $d(A, C) = a + b = 5$ or $d(B, D) = b + c = 5$, contradicting the fact that AC and BD are blue. Similarly, $a + b + c = d$ is possible if (i) $d = 4$ and a, b, c are (in some order) 1, 1, 2 or (ii) $d = 8$ and a, b, c are (in some order) 2, 2, 4. In case (i), either $d(A, C) = 3$ or $d(B, D) = 3$, while in (ii), either $d(A, C) = 6$ or $d(B, D) = 6$, each case leading to a contradiction. The proof that there are no red quadruples is similar.

In fact, Evans, Pulham and Sheehan [143] proved in 1981 that the colouring described above (which is called *the Paley graph of order 17*) is “unique up to relabelling”. In any other red-blue colouring of lines between 17 points (and there are about 2.46×10^{26} such colourings) we can always find either a red or a blue quadruple.

Diagonal Ramsey Numbers and Alien Invasions

In general, Ramsey [314] proved in 1930 that, for every n , there exists an N such that, if N points are connected by red or blue lines, then there exists either n points all connected by red lines, or n points all connected by blue lines. The minimal number N with this property is called *the diagonal Ramsey number* $R(n, n)$. It is trivial that $R(2, 2) = 2$, and we have just proved that $R(3, 3) = 6$, and $R(4, 4) = 18$. One might guess that we can find $R(5, 5)$ by a similar not-so-complicated argument, but in fact determining $R(5, 5)$ remains an open problem despite all efforts, including an extensive computer search. The famous mathematician Paul Erdős said that, if an alien force, much-much more powerful than human civilization, contacted us and said that they will destroy the planet unless we tell them $R(5, 5)$, then we could unite all mathematicians and all computer power in the world to solve the problem. However, if they asked us to determine $R(6, 6)$, we would have a better chance to destroy the aliens...

A Superpolynomial Improvement

Given that the exact computation of $R(n, n)$ is so difficult, can we at least have some estimates? Erdős and Szekeres [137] proved in 1935 that

$$R(n+1, n+1) \leq \frac{(2n)!}{n! \cdot n!}$$

For $R(3, 3)$, this bound gives $R(3, 3) \leq \frac{4!}{2! \cdot 2!} = \frac{1 \cdot 2 \cdot 3 \cdot 4}{1 \cdot 2 \cdot 1 \cdot 2} = 6$, which is the exact value, while for $R(4, 4)$, it gives $R(4, 4) \leq \frac{6!}{3! \cdot 3!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6}{1 \cdot 2 \cdot 3 \cdot 1 \cdot 2 \cdot 3} = 20$, which is close to the correct value $R(4, 4) = 18$. However, as n grows, the gap between the bound and the exact value seems to grow, hence a better bound is desirable.

In 1987, Thomason [375] proved that

$$R(n+1, n+1) \leq n^{-1/2+A/\sqrt{\ln n}} \frac{(2n)!}{n! \cdot n!}$$

for some constant A . For large n , this bound is better than Erdős’ one by a factor of about \sqrt{n} . After 1987, there were no further improvements for more than 20 years, until the following theorem [100] was proved in 2009.

Theorem 9.11 *There exists a constant C such that*

$$R(n+1, n+1) \leq n^{-C \ln n / \ln \ln n} \frac{(2n)!}{n! \cdot n!}$$

No matter what the value of the constant C is, we can find n large enough so that $C \ln n / \ln \ln n$ is larger than, say, 100.5, and, for such values of n , the bound in Theorem 9.11 is better than Thomason's one by a factor about n^{100} , and the same is true if 100 is replaced by any other constant. As mathematicians say, the bound in Theorem 9.11 gives a *superpolynomial improvement* compared to the previous ones.

Because it is known that $\frac{(2n)!}{n! \cdot n!} \leq C \frac{4^n}{\sqrt{n}}$ for some constant C , Erdős' estimate can be rewritten as

$$R(n+1, n+1) \leq C \frac{4^n}{\sqrt{n}},$$

Thomason's theorem implies that

$$R(n+1, n+1) \leq C' \frac{4^n}{n}$$

for some constant C' , while Theorem 9.11 implies that

$$R(n+1, n+1) \leq C_k \frac{4^n}{n^k}$$

for all k , where C_k is a constant depending on k .

Reference

D. Conlon, A new upper bound for diagonal Ramsey numbers, *Annals of Mathematics* **170**-2, (2009), 941–960.

9.12 An Almost Optimal Upper Bound for Moments of the Riemann Zeta Function

A Short Paper of Riemann

In mathematics, seemingly unrelated areas can sometimes become interconnected in an unexpected way. This happened, for example, with number theory and the theory of functions of a complex variable.

In the middle of the 19th century, the mathematician Bernhard Riemann tried to understand a purely number theoretic question: how many prime numbers are there? Prime numbers, those positive integers which have exactly two divisors, 1 and themselves, lie at the heart of number theory. If $\pi(n)$ denotes the number of primes less than or equal to n , there was a conjecture that $\pi(n) \approx \frac{n}{\ln n}$, or, more formally,

that

$$\lim_{n \rightarrow \infty} \frac{\pi(n)}{(n / \ln n)} = 1. \quad (9.16)$$

In 1859, Riemann wrote a short paper [322] on this topic, in which he... did not prove the result. Despite this, the paper had a tremendous influence on the history of mathematics, and, in particular, suggested one of the most famous and important open problems we have ever had.

Functions of a Complex Variable

In this paper, Riemann suggested to attack conjecture (9.16) using methods from a completely different field, the theory of functions of a complex variable. Complex numbers are those of the form $z = a + ib$, with a and b real, where i is an (imaginary) number such that $i^2 = -1$, see e.g. Sect. 1.7 for more details. These numbers were initially invented to solve equations like $x^2 + 1 = 0$, which have no real solutions, but quickly arose in many other applications. Geometrically, a complex number $z = a + ib$ can be represented as a point (a, b) in the coordinate plane. The distance $\sqrt{a^2 + b^2}$ from this point to the coordinate center is called the *absolute value* of z and denoted by $|z|$.

The function $f(z) = z^2$ is an example of a function with a complex argument and complex output, in this case sending the number $z = a + ib$ to the number $(a + ib)^2 = a^2 + 2abi + (ib)^2 = (a^2 - b^2) + (2ab)i$. While $f(z) = z^2$ is defined for all complex numbers, the function $f(z) = 1/z$ is an example of a function defined for all complex numbers except for $z = 0$. The set D_f of all points where f is defined is called the *domain* of f . For any $z_0 \in D_f$, the *derivative* of f at z_0 , denoted by $f'(z_0)$, is defined as

$$f'(z_0) = \lim_{z \rightarrow z_0} \frac{f(z) - f(z_0)}{z - z_0}.$$

This definition is very similar to the definition of the derivative of “usual” functions on the real line, and similar formulas work, e.g. $(z^2)' = 2z$ and $(1/z)' = -1/z^2$, $z \neq 0$. If a function f has a derivative at every point of its domain, it is called *holomorphic* on D_f . For example, $f(z) = z^2$ and $f(z) = 1/z$ are holomorphic functions, while, e.g. the function $f(z) = |z|$ is not, because it has no derivative at $z = 0$.

The Riemann Zeta Function

Before Riemann, it was known that the distribution of the primes depends on the properties of the function

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

where $s > 1$ is a real number, and $\sum_{n=1}^{\infty} \frac{1}{n^s}$ is understood as $\lim_{N \rightarrow \infty} \sum_{n=1}^N \frac{1}{n^s}$. For example, Euler proved in 1734 that $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$, hence $\xi(2) = \frac{\pi^2}{6}$. However, the sum $\sum_{n=1}^{\infty} \frac{1}{n^s}$ is undefined for all real numbers $s < 1$, for example, for $s = 0$ it reduces to $1 + 1 + 1 + \dots$, while for $s = -1$ it reduces to $1 + 2 + 3 + 4 + \dots$.

Riemann noticed that there exists a *unique* function ζ of a *complex* variable which (i) is defined for *all* complex numbers z except $z = 1$, (ii) is holomorphic, and (iii) satisfies $\zeta(z) = \sum_{n=1}^{\infty} \frac{1}{n^z}$ for all z for which the infinite sum is well-defined. This function is called the *Riemann zeta function*. In particular, $\zeta(s)$ is well-defined for real $s < 1$, for example, $\zeta(0) = -\frac{1}{2}$, $\zeta(-1) = -\frac{1}{12}$, which allows mathematician to write various funny formulas like $1 + 1 + 1 + \dots = -\frac{1}{2}$, or $1 + 2 + 3 + 4 + \dots = -\frac{1}{12}$. Also, $\zeta(-2) = \zeta(-4) = \zeta(-6) = \dots = 0$. Numbers of the form $-2k$, $k = 1, 2, 3, \dots$, are called *trivial zeros* of ζ . All other complex numbers z such that $\zeta(z) = 0$ are called *non-trivial zeros*.

The Riemann Hypothesis

Riemann noticed that (9.16) would follow from the following statement

(*) If $z = a + ib$ is a non-trivial zero of ζ , then $a = 1/2$.

The set of all complex numbers $z = 1/2 + ib$, $b \in \mathbb{R}$, is called the *critical line*. Hence, (*) can be reformulated as the conjecture that all non-trivial zeros of ζ lie on the critical line. Figure 9.12a depicts the first few zeros of ζ , and the critical line is drawn as a dotted line. Figure 9.12b depicts the real and imaginary parts of ζ on the critical line, while Fig. 9.12c depicts its absolute value.

At first, (*) looked like a not-very-difficult-to-prove lemma, but Riemann was not able to find a rigorous justification. In 1896, Jacques Hadamard [182] proved a weaker statement “if $z = a + ib$ is a non-trivial zero of ζ , then $a \in [0, 1]$ ”, and was able to deduce (9.16) from it. However, it was clear that even better estimates for the distribution of primes would follow from (*). Statement (*) received the name *Riemann hypothesis* and gained the status of an important open problem. In 1900, Hilbert included it in his list [201] of 23 problems for 20th century mathematics. In 2000, the Clay Mathematics Institute included it in its list of 7 problems, offering a million-dollar prize for its solution. However, the problem is still open, and there is no sign that it will be solved in the near future.

How Large is the Riemann Zeta Function on the Critical Line?

Many other mathematical theorems have been proved in the form “if the Riemann hypothesis holds, then the desired result follows”. However, some other applications require a further understanding of the behaviour of ζ on the critical line $1/2 + it$. In particular, how large is $\zeta(1/2 + it)$? If “large” is understood in terms of absolute

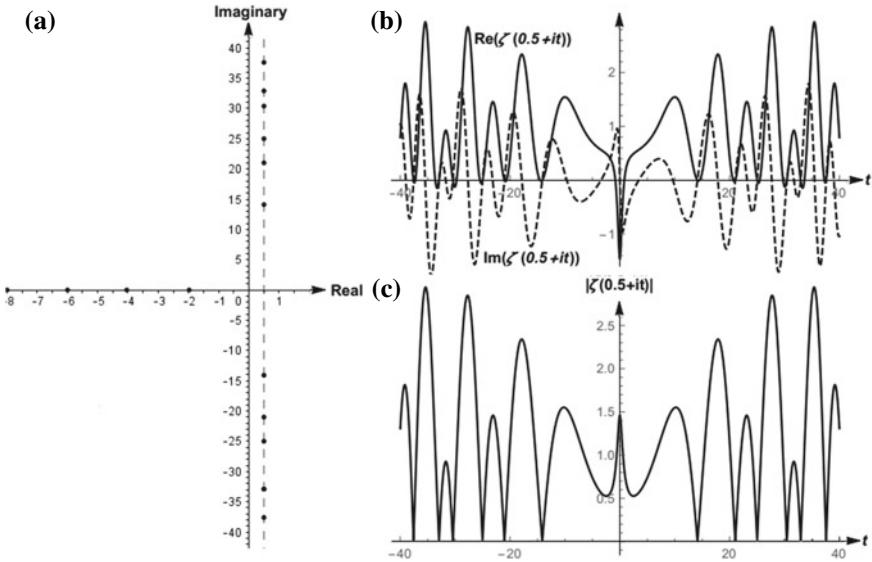


Fig. 9.12 **a** The first few zeros of the Riemann zeta function ζ , **b** and **c** the function ζ on the critical line

value, then the “average size” of $|\zeta(1/2 + it)|$ on the interval $t \in [0, T]$ is given by $\frac{1}{T} \int_0^T |\zeta(1/2 + it)| dt$. One may also be interested in the average size of the squared absolute value $|\zeta(1/2 + it)|^2$, and, more generally, in estimating

$$M_k(T) := \int_0^T |\zeta(1/2 + it)|^{2k} dt,$$

for all $k > 0$. $M_k(T)$ is also called a k -th moment of ζ .

Lower and Upper Bounds for the k -th Moment

Estimating $M_k(T)$ turned out to be a difficult problem, especially if we aim for unconditional results. The progress is better if we assume that the Riemann hypothesis (*) holds. In this case, Ramachandra [313] proved in 1978 that, for every $k > 0$, there is a constant C_k such that

$$C_k T (\ln T)^{k^2} \leq M_k(T), \quad \forall T. \quad (9.17)$$

For an upper bound, the best result (assuming the Riemann hypothesis) before 2009 was

$$M_k(T) \leq C'_k e^{2kC \ln T / \ln \ln T} \quad \forall T,$$

where C'_k is a constant which depends on k , and C is an absolute constant.

The following theorem [354], also assuming the Riemann hypothesis, provides a much better upper bound for all values of k . In fact, the established bound is “ ε -close” to the lower bound of Ramachandra.

Theorem 9.12 *Assume that the Riemann hypothesis (*) holds. Then for every $k > 0$ and every $\varepsilon > 0$, there is a constant $C_{k,\varepsilon}$ such that*

$$M_k(T) \leq C_{k,\varepsilon} T (\ln T)^{k^2+\varepsilon} \quad \forall T.$$

In a later work, Adam Harper [189] improved the bound in Theorem 9.12 and proved that $M_k(T) \leq D_k T (\ln T)^{k^2}$ for some constant D_k . Together with (9.17), this resolves the question of how large $|\zeta|$ is on the critical line, up to a constant factor. The “only” problem is that its resolution is, like hundreds of other important theorems in the field, subject to the correctness of the Riemann hypothesis. If it turns out to be false, the conclusion of all such theorems could be false as well.

Reference

K. Soundararajan, Moments of the Riemann zeta function, *Annals of Mathematics* **170**-2, (2009), 981–993.

9.13 Optimal Lattice Sphere Packing in Dimension 24

Vectors and Lattices

Vectors in the plane are, geometrically, directed line segments, connecting an initial point A with a terminal point B , and usually denoted \mathbf{AB} . In the coordinate plane, we say that \mathbf{AB} has coordinates $(x_B - x_A, y_B - y_A)$, where (x_A, y_A) and (x_B, y_B) are coordinates of A and B , respectively. In particular, if $O = (0, 0)$ is the coordinate center, then \mathbf{OA} has the same coordinates as A . Vectors can be added and multiplied by constants using rules $(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$ and $\lambda(x, y) = (\lambda x, \lambda y)$.

A lattice $\mathcal{L} = \mathcal{L}_{A,B}$ in the plane is the set of points X such that $\mathbf{OX} = k\mathbf{OA} + m\mathbf{OB}$, where A and B are some fixed points such that O, A and B are not on the same line, and k, m are integers. For example, if A and B have coordinates $(1, 0)$ and $(0, 1)$, respectively, then $k(1, 0) + m(0, 1) = (k, m)$, hence the lattice $\mathcal{L}_{A,B}$ consists of all points with integer coefficients.

Counting Lattice Points per Unit Area

If we draw a circle with center $O = (0, 0)$ and large radius R , how many points of $\mathcal{L}_{A,B}$ does it contain? To estimate this number, which we denote by $N(\mathcal{L}_{A,B}, R)$, let us associate to every lattice point $X = (k, m)$ the unit square U_X for which X is the

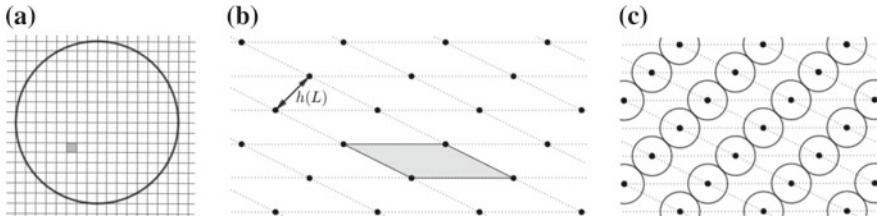


Fig. 9.13 Lattices, the fundamental parallelogram, and sphere packing

left bottom vertex, that is, the square with vertex coordinates $(k, m), (k + 1, m), (k + 1, m + 1), (k, m + 1)$, see Fig. 9.13a. In most cases, a point X is inside the circle if and only if U_X is inside it as well. This is not true if X is near the boundary of the circle, but, for large R , the number of lattice points near the boundary is much less than $N(\mathcal{L}_{A,B}, R)$, and this boundary effect can be ignored. Now, the number of unit squares inside the circle is, again up to boundary effects, equal to the ratio $S(\text{Circle})/S(U_X)$, where $S(\text{Circle}) = \pi R^2$ and $S(U_X)$ are the areas bounded by the circle and U_X , respectively. Hence, the circle contains roughly about $\pi R^2/S(U_X)$ lattice points, or, in other words, about $1/S(U_X)$ lattice points on average per unit area. The quantity $1/S(U_X)$ is called the *density* of the lattice \mathcal{L} in the plane. In our case, U_X is a unit square, $S(U_X) = 1$, hence its density $1/S(U_X)$ is also equal to 1: our lattice contains on average 1 point per unit area.

The Fundamental Parallelogram, and the Density of a Lattice

If $A = (1/2, 0)$, $B = (0, 1/2)$, the lattice $\mathcal{L}_{A,B}$ consists of the points whose coordinates are either integers or half integers. In this case, every point $X \in \mathcal{L}_{A,B}$ is a left bottom vertex of a square U_X with side length 0.5. Hence, $S(U_X) = 0.25$, and the density $1/S(U_X)$ is equal to 4, that is, there are four points of the lattice per unit area. A slightly more complicated example is $A = (3, 0)$, $B = (-2, 1)$. Then $\mathcal{L}_{A,B}$ consists of points with coordinates $k(3, 0) + m(-2, 1) = (3k - 2m, m) = (3[k - m] + m, m)$, that is, of all points X with integer coordinates (x, y) , such that $x - y$ is a multiple of 3. In this case, X is a left bottom vertex of the parallelogram U_X with vertex coordinates $(x, y), (x + 3, y), (x + 1, y + 1), (x - 2, y + 1)$, see Fig. 9.13b. The area $S(U_X)$ and density $1/S(U_X)$ are then equal to 3 and $1/3$, respectively.

In general, the *fundamental parallelogram* U of a lattice $\mathcal{L}_{A,B}$ in the plane is the set of points X such that $\mathbf{OX} = \alpha\mathbf{OA} + \beta\mathbf{OB}$, where $\alpha \in [0, 1]$ and $\beta \in [0, 1]$. In other words, U is the parallelogram with vertices O, A, C, B , where C has coordinates $(x_A + x_B, y_A + y_B)$. The whole lattice can be considered as the vertices of a tiling of the plane by copies of this parallelogram. The real number $\rho(\mathcal{L}_{A,B}) := 1/S(U)$ is called the *density* of $\mathcal{L}_{A,B}$.

The Length of the Shortest Vector and Circle Packing

Another important parameter of any lattice \mathcal{L} is the length of the shortest vector in it

$$h(\mathcal{L}) := \min_{X \in \mathcal{L}} |\mathbf{OX}| = \min_{k, m \in \mathbb{Z}} |k\mathbf{OA} + m\mathbf{OB}|,$$

where $|\cdot|$ denotes the length of a vector, \mathbb{Z} denotes the set of all integers, and the minimum is with respect to all possible integer values of k, m except for $k = m = 0$. For example, for the lattice $\mathcal{L}_{A,B}$ defined by $A = (1, 0)$ and $B = (0, 1)$, we have $h(\mathcal{L}_{A,B}) = \min_{k,m} \sqrt{k^2 + m^2} = 1$. Similarly, if $A = (1/2, 0)$, $B = (0, 1/2)$, then $h(\mathcal{L}_{A,B}) = 1/2$, while in the lattice with $A = (3, 0)$, $B = (-2, 1)$, $h(\mathcal{L}_{A,B}) = \min_{k,m} \sqrt{(3k - 2m)^2 + m^2} = \sqrt{2}$, with the minimum achieved, for example, for $k = m = 1$.

Geometrically, $h(\mathcal{L})$ is the minimal distance from the coordinate center $O = (0, 0)$ to any other point of the lattice. In fact, $h(\mathcal{L})$ is also the minimal distance between *any* two points of the lattice, see Fig. 9.13b. Indeed, if $\mathbf{OX} = k_1\mathbf{OA} + m_1\mathbf{OB}$ and $\mathbf{OY} = k_2\mathbf{OA} + m_2\mathbf{OB}$, then

$$|\mathbf{XY}| = |\mathbf{OY} - \mathbf{OX}| = |(k_2 - k_1)\mathbf{OA} + (m_2 - m_1)\mathbf{OB}| \geq h(\mathcal{L})$$

provided that $X \neq Y$. In particular, this implies that if we draw circles of radii $h(\mathcal{L})/2$ with centres at every point of \mathcal{L} , then these circles would not intersect, see Fig. 9.13c. In other words, we can draw $\rho(\mathcal{L})$ non-intersecting circles (per unit area) of radii $h(\mathcal{L})/2$, or, equivalently, $\rho(\mathcal{L})(h(\mathcal{L})/2)^2 = \frac{h^2(\mathcal{L})}{4S(U)}$ non-intersecting circles (per unit area) of radii 1.

This motivates the search for lattices \mathcal{L} with ratio

$$r(\mathcal{L}) := \frac{h^2(\mathcal{L})}{4S(U)}$$

as large as possible. For example, $r(\mathcal{L}_{A,B}) = 1^2/(4 \cdot 1) = 1/4$ if $A = (1, 0)$ and $B = (0, 1)$, while $r(\mathcal{L}_{A,B}) = (\sqrt{2})^2/(4 \cdot 3) = 1/6$. It turns out that $r(\mathcal{L})$ is maximal if the points O, A, B form an equilateral triangle, that is, $A = (1, 0)$, $B = (1/2, \sqrt{3}/2)$. Then $S(U) = \sqrt{3}/2$, $h(\mathcal{L}) = 1$, and $r(\mathcal{L}) = 1^2/(4 \cdot \sqrt{3}/2) = 1/2\sqrt{3} \approx 0.29$.

Lattice-Based Sphere Packings in Higher Dimensions

The same question can be asked in any dimension. In dimension n , points and vectors are described by n coordinates, (x_1, x_2, \dots, x_n) , e.g. the coordinate center O is $(0, 0, \dots, 0)$. The length $|\mathbf{a}|$ of a vector $\mathbf{a} = (x_1, x_2, \dots, x_n)$ is $|\mathbf{a}| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$. Any n vectors $\mathbf{a}_1, \dots, \mathbf{a}_n$ define a lattice $\mathcal{L} = \mathcal{L}(\mathbf{a}_1, \dots, \mathbf{a}_n)$, which is the set of all points X such that $\mathbf{OX} = k_1\mathbf{a}_1 + k_2\mathbf{a}_2 + \dots + k_n\mathbf{a}_n$ for some inte-

gers k_1, k_2, \dots, k_n . The *fundamental parallelepiped* $U_{\mathcal{L}}$ of a lattice \mathcal{L} is the set of points X such that $\mathbf{OX} = \beta_1 \mathbf{a}_1 + \beta_2 \mathbf{a}_2 + \dots + \beta_n \mathbf{a}_n$, where each β_i is a real number such that $0 \leq \beta_i \leq 1, i = 1, 2, \dots, n$. If the volume $V(U_{\mathcal{L}})$ of $U_{\mathcal{L}}$ in n -dimensional space is non-zero, $1/V(U_{\mathcal{L}})$ has the meaning of the average number of points in \mathcal{L} per unit volume.

The length $h(\mathcal{L})$ of the shortest vector in \mathcal{L} is

$$h(\mathcal{L}) := \min_{X \in \mathcal{L}} |\mathbf{OX}| = \min_{(k_1, k_2, \dots, k_n) \in \mathbb{Z}^n / 0} |k_1 \mathbf{a}_1 + k_2 \mathbf{a}_2 + \dots + k_n \mathbf{a}_n|,$$

where the notation $\mathbb{Z}^n / 0$ means that the minimum is with respect to all possible integer values of k_1, k_2, \dots, k_n except for $k_1 = k_2 = \dots = k_n = 0$. Using points in \mathcal{L} as centres, we can locate $1/V(U_{\mathcal{L}})$ n -dimensional spheres per unit volume, each of radius $h(\mathcal{L})/2$, such that the interiors of the spheres do not intersect. After scaling, this allows us to locate $(h(\mathcal{L})/2)^n / V(U_{\mathcal{L}})$ non-intersecting spheres (per unit volume) of radius 1 each. This motivates the question of finding, for each dimension n , a lattice \mathcal{L} with ratio

$$r(\mathcal{L}) := \frac{h^n(\mathcal{L})}{2^n V(U_{\mathcal{L}})}$$

as large as possible. For $n = 3$ (the “usual” three-dimensional space we live in) this question was resolved by Gauss in 1831. Korkine and Zolotareff [228, 229] resolved the $n = 4$ case in 1873, and the $n = 5$ case in 1877, while Blichfeldt [59] resolved the cases $6 \leq n \leq 8$ in 1935. However, no further case had been solved for 74 years, until the following theorem was proved in [95].

Theorem 9.13 *In dimension $n = 24$, the maximal possible value of $r(\mathcal{L})$ is equal to 1.*

It is interesting that, after dimensions $1 \leq n \leq 8$, the next resolved case is $n = 24$. The 24-dimensional lattice \mathcal{L} with $r(\mathcal{L}) = 1$ was found by Leech in 1964, and has the name Leech lattice. The contribution of Theorem 9.13 is the proof that no 24-dimensional lattice \mathcal{L} has $r(\mathcal{L}) > 1$. Moreover, the authors also proved that the Leech lattice is the only one with $r(\mathcal{L}) = 1$, up to scaling and isometries.

Theorem 9.13 implies that sphere packing using the Leech lattice is the densest possible one among all lattice-based sphere packings in dimension 24. In a later work [96], Cohn, Kumar, Miller, Radchenko, and Viazovska proved that this sphere packing is in fact the densest possible one among all packings in dimension 24, not necessary lattice-based ones.

Reference

H. Cohn and A. Kumar, Optimality and uniqueness of the Leech lattice among lattices, *Annals of Mathematics* **170**-3, (2009), 1003–1050.

9.14 A Waring-Type Theorem for Large Finite Simple Groups

Representing Integers as Sums of Perfect Powers

One of the oldest classical topics in mathematics is representing integers as a sum of some “special” integers. For example, the Greek mathematician Diophantus, who lived in the 3rd century, was interested in representing integers as a sum of perfect squares, e.g. $1 = 1^2$, $2 = 1^2 + 1^2$, $3 = 1^2 + 1^2 + 1^2$, $4 = 2^2$, $5 = 2^2 + 1^2$, $6 = 2^2 + 1^2 + 1^2$, $7 = 2^2 + 1^2 + 1^2 + 1^2$, and so on. You can see that we need at least 4 squares to represent 7. Diophantus was interested in the question of whether there exists a positive integer which requires at least 5 squares for such a representation, or if 4 squares always suffice. This question was answered by Lagrange in 1770. His celebrated four squares theorem states that four squares suffice: every positive integer n can be written as $n = a^2 + b^2 + c^2 + d^2$ for some integers a, b, c, d .

In the same year as Lagrange proved his theorem, Edward Waring asked if similar results can be proved for cubes, fourth powers, and so on. That is, does there exist a positive integer N_3 such that every positive integer n can be written as a sum of at most N_3 cubes? More generally, for every k , does there exist an N_k such that every positive integer n can be written as a sum of at most N_k k -th powers? This question was answered positively by Hilbert [202] in 1909, and is known as the Hilbert–Waring theorem.

What is the Minimal Number of k -th Powers We Will Need?

It follows from Lagrange’s theorem that the statement “every positive integer n can be written as a sum of at most N_2 squares” holds with $N_2 = 4$, and the example of $n = 7$ shows that it does not hold for $N_2 = 3$. In other words, 4 is the *minimal* number of squares sufficient to represent every integer. One may then ask for the *minimal* number of cubes, 4th powers, and so on. In general, let $g(k)$ be the *minimal* number of k -th powers sufficient to represent every positive integer.

By 1912, Wieferich and Kempner [217, 403] had shown that every integer is the sum of 9 cubes. Because 23 cannot be represented as a sum of 8 cubes, this proves that $g(3) = 9$. Later, mathematicians proved that $g(4) = 19$, $g(5) = 37$, and so on. In fact, it is now known that $g(k) = 2^k + [(3/2)^k] - 2$ for all values of k , except for possibly finitely many exceptions. Here, $[(3/2)^k]$ denotes the largest integer less than $(3/2)^k$.

While the representation of 23 requires 9 cubes, Linnik [246] proved in 1943 that all $n > 454$ can be represented as a sum of at most 7 cubes. Also, while $g(4) = 19$, it is known that all $n > 13792$ are the sums of only 16 4th powers. The question “for given k , what is the minimal number of k -th powers required to represent any sufficiently large n ” remains an active area of research today.

Representing Rotations as a Composition of Some “Special” Rotations

Similar questions of the form “Can we represent an object using some “special” objects?” can be asked not only about integers, but in many areas of mathematics. In geometry, one may study rotations of the plane around some fixed center O by some arbitrary angle α . If we perform any such rotation R' , and then another rotation R'' , the result is again a rotation, which we denote by $R'' \circ R'$, and call the *composition* of R' and R'' . Let S be a set of rotations for which α has n degrees for some integer n . Let us call “perfect squares” some special rotations from S , which can be written as $R \circ R$ for some $R \in S$. For example, if R_1 is a rotation clockwise by angle 1° , then $R_2 = R_1 \circ R_1$ is a rotation clockwise by angle 2° , and, by definition, R_2 is a perfect square. Now, by analogy with Lagrange’s theorem, one may ask if any rotation $R \in S$ can be represented as a composition of such “perfect squares”. In fact, the answer is “no”. One can easily check that all “perfect squares” rotate the plane by an *even* number of degrees, and so do their compositions, hence any rotation by an odd number of degrees, such as R_1 , cannot be represented as such a composition.

Representing Permutations as a Composition of Some “Special” Permutations

As another example, let us consider functions from some *finite* set S to itself. If S has n elements, we can enumerate them, and write S as $\{1, 2, \dots, n\}$. Then any function $f : S \rightarrow S$ can be described by listing its values: $f = (f(1), f(2), \dots, f(n))$. For example, if $n = 3$, $S = \{1, 2, 3\}$, then the function $f(x) = 1$ (constant function) is written as $(1, 1, 1)$, while the function $f(x) = x$ is written as $(1, 2, 3)$. If all $f(i)$, $i = 1, 2, \dots, n$, are different, then f is called a *permutation*. For example, $(1, 1, 1)$ is not a permutation, while $(1, 2, 3)$ is. In general, let S_n be the set of all permutations $g : \{1, 2, \dots, n\} \rightarrow \{1, 2, \dots, n\}$.

The composition $g \circ f$ of any functions f and g is the function h such that $h(x) = g(f(x))$ for all x , and one can easily prove that the composition of any two permutations is again a permutation. Let us call a function $g \in S_n$ a “perfect square” if $g = f \circ f$ for some $f \in S_n$. Can any $h \in S_n$ be written as a composition of perfect squares? It turns out, not. For $n = 3$, there are exactly 6 permutations: $a = (2, 1, 3)$, $b = (1, 3, 2)$, $c = (2, 3, 1)$, $d = (3, 1, 2)$, $e = (1, 2, 3)$, and $f = (3, 2, 1)$. One can check that $a \circ a = b \circ b = e \circ e = f \circ f = e$ while $c \circ c = d$ and $d \circ d = c$, see Fig. 9.14, hence the perfect squares are e, c and d . Next, $e \circ c = c \circ e = c$, $e \circ d = d \circ e = d$, and $c \circ d = d \circ c = e$, hence the composition of any two perfect squares is again a perfect square, and any permutation outside the set $\{e, c, d\}$ cannot be represented in this way.

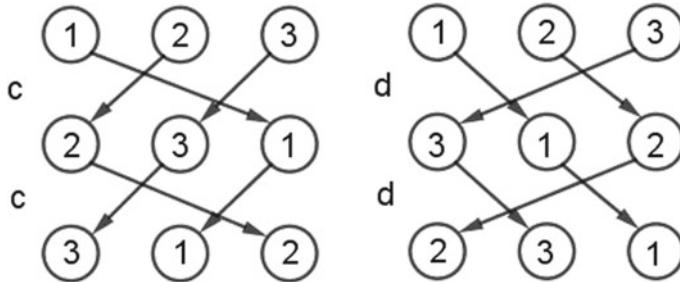


Fig. 9.14 Illustration for $c \circ c = d$ and $d \circ d = c$ in S_3

Even Permutations

In general, a permutation $f \in S_n$ is called *even* if the number of pairs (i, j) such that $i < j$ but $f(i) > f(j)$ is even. In other words, a permutation is even if it exchanges the order of an even number of pairs (i, j) . For example, the permutation d in Fig. 9.14, sending $(1, 2, 3)$ to $(3, 1, 2)$, exchanges the order in the pair $(2, 3)$ (2 was on the left of 3 before permutation, but on the right of 3 after permutation), and in the pair $(1, 3)$, but does not change the order in the pair $(1, 2)$ (1 was on the left of 2 before permutation, and stays on the left of 2 after permutation). Hence, the total number of pairs with exchanged order is 2, an even number, and this permutation is an even permutation.

The set of all even permutations is usually denoted by A_n . One can prove that all perfect squares always belong to A_n , and so do all their compositions. Hence, there is no hope of representing every permutation $f \in S_n$ as a composition of perfect squares. However, one may ask if at least every even permutation $f \in A_n$ is representable in this way, and if so, how many perfect squares we would need for such a representation. The same question can be asked for cubes, and, more generally, for k -th powers for arbitrary k .

Groups, Subgroups, and Simple Groups

In the above examples, we considered integers, rotations of the plane, and permutations. All these are example of *groups*. A group is an arbitrary set G together with an operation \cdot such that (i) $a \cdot b \in G$ for all $a, b \in G$; (ii) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all $a, b, c \in G$; (iii) there exists an $e \in G$ (called the identity element of G) such that $a \cdot e = e \cdot a = a$ for all $a \in G$; and (iv) for every $a \in G$, there exists an element $a^{-1} \in G$ (called the *inverse* of a), such that $a \cdot a^{-1} = a^{-1} \cdot a = e$. The set of integers form a group (usually denoted by \mathbb{Z}) with the addition operation $+$, while rotations and permutations form a group with the composition operation \circ .

A subset H of a group G is called a *subgroup* of G if (i) $a \cdot b \in H$ for all $a, b \in H$; (ii) $e \in H$, where e is the identity element of G ; and (iii) $a^{-1} \in H$ for every $a \in H$.

For example, the set of all even integers is a subgroup of \mathbb{Z} , while A_n is a subgroup of S_n . A subgroup H of a group G is called *trivial* if either $H = G$, or $H = \{e\}$, and *non-trivial* otherwise. A subgroup H of a group G is called *normal* if $g \cdot a \cdot g^{-1} \in H$ for any $a \in H$ and $g \in G$. One can check that A_n is a normal subgroup of S_n .

A group G is called *simple* if it does not have any non-trivial normal subgroups. For example, the group S_n is not simple, because it has a non-trivial normal subgroup A_n . However, it turns out that, for $n \geq 5$, A_n has no non-trivial normal subgroups, hence it is a simple group.

Representing Group Elements as a Composition of Some “Special” Elements

A *square* in a group G is any element $a \in G$ which can be written as $a = b \cdot b$ for some $b \in G$. Similarly, $a \in G$ is called a *k -th power*, if $a = b \cdot b \cdots b$ (k times) for some $b \in G$. One may ask if every element $a \in G$ can be written as a composition of squares, or, more generally, k -th powers. In general, the answer is “no”, because all squares (or k -th powers) can belong to some normal subgroup H of G , and so are all their compositions. This motivates us to study the same question for the case when G is a *simple* group. It turns out that in this case the answer is “yes”, and one may then look for a *minimal* number of squares (or k -th powers) needed for such a representation.

In fact, squares and k -th powers are just special cases of the general notion of group words. A *word* w is any finite string of symbols, possibly with repetitions and with inverse symbol, like aaa or $aabbcc^{-1}a^{-1}a^{-1}c$. Let w be a word with d different symbols s_1, s_2, \dots, s_d , G be a group, and $g_1, g_2, \dots, g_d \in G$ be any d elements of G . Then we write $w(g_1, g_2, \dots, g_d)$ to be the result of (i) substitution of g_1, g_2, \dots, g_d into w instead of s_1, s_2, \dots, s_d , respectively, and (ii) performing the group operation. Let $w(G)$ denote the set of all elements $g \in G$ representable in the form $g = w(g_1, g_2, \dots, g_d)$ for some $g_1, g_2, \dots, g_d \in G$. For example, the set of all squares in G is just $w(G)$ for $w = aa$, while the set of all k -th powers is $w(G)$ for $w = aa \dots a$ (k times).

One can then ask if every element of a simple group G can be represented as a composition of elements from $w(G)$, and if so, how many elements from $w(G)$ we need for this. The following theorem, proved in [345], states that, for sufficiently large (but finite) G , every $g \in G$ is in fact a composition of just *three* elements from $w(G)$!

Theorem 9.14 *Let w be any non-empty word. Then there exists a positive integer N , depending only on w , such that for every finite simple group G with $|G| \geq N$, every element $g \in G$ can be represented as $g = g_1 \cdot g_2 \cdot g_3$, where $g_i \in w(G)$, $i = 1, 2, 3$.*

Reference

- A. Shalev, Word maps, conjugacy classes, and a noncommutative Waring-type theorem, *Annals of Mathematics* **170**-3, (2009), 1383–1416.

Chapter 10

Theorems of 2010



10.1 Majority Votes Are the Most Stable

Some Difficulties in Roommate Selection

Joanna, an undergraduate who is looking for a roommate to share rent, evaluates potential roommates based on three criteria: sense of humour, intelligence, and character. For each criterion, she assigns a score 0, 1, or 2, see Table 10.1, and prefers one roommate to another if she is better in at least two out of three criteria. She recently rejected a roommate offer from Christina, evaluated (2, 1, 0) (a funny girl with a great sense of humour, reasonably intelligent, but not of very good character) in favour of Veronica, evaluated (0, 2, 1), because she is more intelligent and of better character. Later, however, she decided to choose Victoria, evaluated (1, 0, 2), because she has even better character, and is also funnier. After this, Joanna regretted rejecting Christina, because, comparing to Victoria, she is funnier and more intelligent (Fig. 10.1).

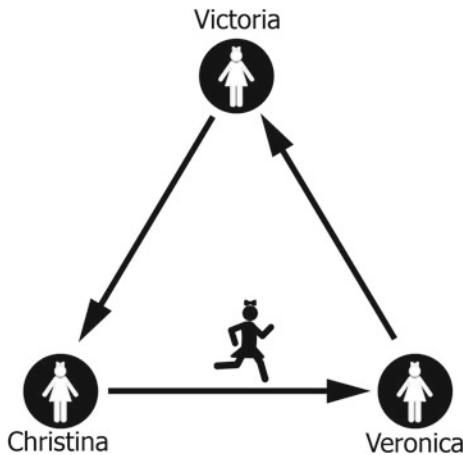
Nine People and Three Alternatives

A similar paradox may occur in the process of voting. Assume that 9 people need to select one out of three possible alternatives A , B , C . They may be workers in some company who need to select a boss out of three candidates, or investors deciding in which out of three projects to invest their joint capital, or even a family deciding among three variants where to go for a holiday together. Further assume that everyone has a strong opinion which he/she is not going to change. These opinions are presented in Table 10.2.

Because they cannot convince each other, it seems that the only reasonable option for them is to vote, and select an alternative which has the most votes. However, how would you organize the voting?

Table 10.1 Joanna's criteria for roommate selection

	Sense of humour	Intelligence	Character
Excellent (score 2)	Christina	Veronica	Victoria
Good (score 1)	Victoria	Christina	Veronica
Average (score 0)	Veronica	Victoria	Christina

**Fig. 10.1** Difficulties in roommate selection**Table 10.2** Preferences of 9 people among 3 alternatives

	1	2	3	4	5	6	7	8	9
1st choice	A	A	A	A	B	B	B	C	C
2nd choice	C	C	C	C	A	C	C	B	B
3rd choice	B	B	B	B	C	A	A	A	A

Three Natural Voting Systems Lead to Three Different Winners

Supporters of alternative *A* suggest that everyone should name their favourite alternative, and the alternative with most votes wins. In this case, *A* would get 4 votes, *B* would get 3, and *C* would get 2, so it seems that the group should select *A*.

However, supporters of *B* disagree with such a voting system. Why should we agree with alternative *A*, they say, on the basis that it has got 4 votes out of 9, that is, 4 people like it and 5 dislike it? Why should the majority agree with the minority? They insist that neither alternative can win unless it has got the *majority* of votes, and suggest the two-round system which is used for presidential elections in many

countries: if no alternative gets the majority of votes during the first round, then two “leaders” go to the second round. In our case, they say, $A \rightarrow 4$, $B \rightarrow 3$, $C \rightarrow 2$ is the result of the first round, hence A and B should go to the second round, in which supporters of C (people 8 and 9 in the table) would vote for B , so B would get the majority of votes (5 out of 9) and should be a winner.

Wait a minute, say supporters of C . It is unfair to base the voting on the first choices only. Let us award to every alternative 2 points for every person who selects it as a first choice, 1 point for every 2nd choice, and 0 points for the 3rd choice. Similar voting systems are used everywhere, for example, in the selection of the best player of the year (the Golden Ball award) in football, or the best singer in the Eurovision competition. In this case, A gets 4 first choices and 1 second choice, $4 \cdot 2 + 1 = 9$ points in total, B gets 3 first choices and 2 second choices, $3 \cdot 2 + 2 \cdot 1 = 8$ points, while C gets 2 first choices and 6 second choices, $2 \cdot 2 + 6 \cdot 1 = 10$ points. Hence, alternative C should be the winner.

Even if the group agrees on some voting system, and selects some alternative, they will have a similar problem to Joanna’s one. If they select A , then the majority (people 5, 6, 7, 8, 9 in the table) would be willing to switch to B . If they select B , then 6 people out of 9 (1, 2, 3, 4, 8, 9 in the table) would want to switch to C . However, with choice C , the majority (people 1, 2, 3, 4, 5 in the table) will regret not choosing A .

General Voting Rules

In our example, whatever alternative people choose, most of them will be unhappy and will insist on changing it to another one. Similarly, whatever roommate Joanna selects, she would regret not choosing another one. We will call this the *pairwise comparison paradox*. These examples suggest that the majority rule may not be an ideal choice for the pairwise comparison of the alternatives. Do there exist other selection rules (instead of the majority rule) which would help to avoid this paradox? To answer this question, we first need to formally define what exactly we mean by a “selection rule” or “voting rule” between two alternatives. For convenience, we will name two alternatives as “1” and “−1”. If n people will vote, let us denote by x_i the result of the i -th vote, that is, $x_i = 1$ if the i -th vote is for alternative 1 and $x_i = -1$ otherwise. Similarly, if n criteria are considered, then $x_i = 1$ if alternative 1 is better with respect to criterion i and $x_i = -1$ otherwise. Then the result of any voting/comparison is just a vector $x = (x_1, x_2, \dots, x_n)$. The “voting rule” is any function f which assigns to each such vector x the value 1 or −1, which represents the “winner”. In this notation, the majority rule is the function

$$f_M(x) = \operatorname{sgn} \left(\sum_{i=1}^n x_i \right),$$

where $\text{sgn}(z) = 1$ if $z > 0$ and $\text{sgn}(z) = -1$ if $z < 0$. For example, in the Christina-Veronica comparison in Table 10.1, denote Christina as 1, and Veronica as -1 , then $x_{CV} = (1, -1, -1)$, and the majority rule returns $f_M(x_{CV}) = \text{sgn}(1 - 1 - 1) = -1$, which means that Veronica is preferred to Christina.

Another example of a voting rule is $f_1(x) = f_1(x_1, x_2, \dots, x_n) = x_1$, that is, just accept the choice of the first voter, or use the first criterion. For example, $f_1(x_{CV}) = f_1(1, -1, -1) = 1$, so, with this rule, Christina is preferred to Veronica. Of course, we may design rules $f_i(x) = f_i(x_1, x_2, \dots, x_n) = x_i$ for every $i = 1, 2, \dots, n$. These rules are called “dictatorship” rules, because they suggest following votes/criterion i , ignoring everything else.

Interestingly, dictatorship rules help us to avoid the “paradox of pairwise comparison” described above. For example, if Joanna decides to use rule f_1 , she will select the roommate based on humour sense, and, in our example, would stay with Christina. With dictatorship rule f_2 , she would select Veronica because of her intelligence, and never regret it.

Voting Rules with Low Influences

Of course, dictatorship rules are far from ideal. Joanna may find it unreasonable to select a roommate based on one criterion only, ignoring everything else. The use of dictatorship rules in voting systems looks even less reasonable. In the example presented in Table 10.2, dictatorship rule f_i implies that everyone should accept the alternative which person i prefers, ignoring the opinions of everyone else. Intuitively, voting rule f is “good” if each voter/criterion i has only limited influence on the final decision. Formally, we define the *influence of i on f* by

$$\text{Inf}_i(f) := \mathbb{P}[f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) \neq f(x_1, \dots, x_{i-1}, -x_i, x_{i+1}, \dots, x_n)],$$

where \mathbb{P} denotes the probability, assuming that each x_j , $j \neq i$, is selected independently at random, with x_j being equal to -1 or 1 with equal chance. In other words, $\text{Inf}_i(f)$ is the chance that person i will change the voting result, if everyone else votes at random. For example, with $n = 3$ and the majority rule, the vote of the first person can change the final result if $x_2 = 1, x_3 = -1$ or if $x_2 = -1, x_3 = 1$, but cannot change anything if $x_2 = x_3 = 1$ or $x_2 = x_3 = -1$, hence $\text{Inf}_1(f_M) = \frac{2}{4} = 0.5$. For dictatorship rules, $\text{Inf}_1(f_1) = 1$ but $\text{Inf}_1(f_2) = \text{Inf}_1(f_3) = 0$, that is, the first person has influence 1 if he/she is a dictator, but influence 0 if the dictator is someone else. Ideally, we would like to have a voting system f with limited influence of each voter, that is, we would like the inequality $\text{Inf}_i(f) \leq \delta$, $\forall i$ to hold for some $\delta \in (0, 1)$, the smaller the better. Another natural requirement for the voting rule f is that $E[f] = 0$, where $E[f]$ is the average value of f over all 2^n inputs. This condition implies that voting rule f gives no preferences to any alternative.

What is the Chance of Avoiding the Paradox?

The famous Arrow's Impossibility Theorem [15] implies that, for any non-dictatorship voting rule f , we can design a situation like in Tables 10.1 and 10.2 with three alternatives A, B, C , such that pairwise comparison using rule f leads to “cyclical” preferences, say C is better than B , which is better than A , which is better than C . However, could it be that such examples are artificial, and have little chance of occurring in practice? In 2002, Kalai [214] studied the probability that such a situation will *not* happen, provided that the preferences of each voter are selected randomly among the 6 possible rankings of A, B, C . He found that this probability is equal to $\frac{3}{4}(1 + S_{1/3}(f))$, where $S_\rho(f)$, $0 \leq \rho \leq 1$, is called the *noise stability* of f , and is defined as $S_\rho(f) := E[f(x)f(y)]$, where $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ are selected at random in such a way that, for all $i = 1, 2, \dots, n$, $\mathbb{P}[x_i = 1] = \mathbb{P}[x_i = -1] = \mathbb{P}[y_i = 1] = \mathbb{P}[y_i = -1] = 1/2$, and $\mathbb{P}[x_i = y_i = 1] = (1 + \rho)/4$. Of course, the higher $S_\rho(f)$ for $\rho = 1/3$, the higher the chance $\frac{3}{4}(1 + S_{1/3}(f))$ of avoiding the paradox. Motivated by this, it is natural to look for a voting rule f with $E[f] = 0$ and low influence $\text{Inf}_i(f)$ of each index i , but with as high stability $S_\rho(f)$ as possible. In fact, Kalai conjectured that (under some additional conditions and for $\rho = 1/3$), the best such f is... the majority rule! This statement later became known as the “Majority is stablest” conjecture.

Majority is Stablest and Approximating Max-Cut

In 2007, Khot, Kindler, Mossel, and O'Donnell [225] studied a seemingly different problem. Given a graph (that is, a set of vertices, some of which are connected by edges), the task is to divide its vertices in two groups, say A and B , such that the number of edges with endpoints in different groups is as large as possible (this problem is called Max-Cut). Of course, one may just try all possible groups A and B , but, if the graph has n vertices, there are 2^n ways to divide them into two groups, and, for reasonably large n (say $n = 100$), trying them all would take too much time even on the fastest computers. Hence, people are interested in looking for more efficient algorithms to solve such problems, with special attention to polynomial algorithms, that is, those which can finish the task after at most $P(n)$ operations for some polynomial P . For Max-Cut, no polynomial algorithm is known to solve the problem exactly, but there is one which can find the solution approximately, to within a factor of about 0.87856. Khot et al were trying to prove that this is optimal, and no polynomial algorithm can find a better approximation. They succeeded in proving their result conditionally, assuming several unproven conjectures. One of the conjectures they need is (the general form of) the “Majority is stablest” conjecture. At first glance, it is surprising that a theorem about algorithms on graphs is connected to a conjecture about voting rules, but this is indeed the case.

The Proof

It often happens in mathematics that people prove some theorems assuming a conjecture, but cannot prove the conjecture itself. The conjecture may then remain open for centuries, or even turn out to be false, which invalidates the theorems. Sometimes, however, we have a happier story when the conjecture turns out to be correct, and the proof follows reasonably soon. This is what happened with the “Majority is stablest” conjecture: from 2010, it became a theorem [283].

Theorem 10.1 *Let $0 \leq \rho \leq 1$ and $\varepsilon > 0$ be given. Then there exists a $\delta > 0$ such that if voting rule f satisfies $E[f] = 0$ and $\text{Inf}_i(f) \leq \delta$, $\forall i$, then*

$$S_\rho(f) \leq \varepsilon + \lim_{n \rightarrow \infty} S_\rho(f_{M_n}),$$

where f_{M_n} is the majority rule for n voters. In other words, the “majority is stablest” conjecture is true.

In fact, the authors of [283] developed a much more general theory, applicable to a more general class of functions (not necessary voting rules). Theorem 10.1, together with some other interesting results, is just a corollary of this theory.

Reference

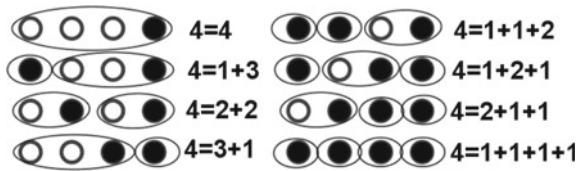
E. Mossel, R. O’Donnell, K. Oleszkiewicz, Noise stability of functions with low influences: Invariance and optimality, *Annals of Mathematics* **171**-1, (2010), 295–341.

10.2 On Divisibility Properties of Dyson’s Rank Partition Function

Representing a Positive Integer as a Sum of Other Positive Integers

How many ways are there to write a positive integer n as a sum of other positive integers? Equivalently, how many ways are there to divide n identical objects into boxes? For example, $n = 2$ objects can either be put into the same box ($2 = 2$), or into different boxes ($2 = 1 + 1$), hence there are 2 ways in total. Three objects can be in the same box ($3 = 3$), in two boxes ($3 = 2 + 1 = 1 + 2$), or in three boxes ($3 = 1 + 1 + 1$). The question is now whether we should count $3 = 2 + 1$ and $3 = 1 + 2$ as different ways. If we count them as different ways, then there are 4 ways to represent 3, and, in general, it is not difficult to prove that there are 2^{n-1} ways to represent n . Indeed, given a representation $n = k_1 + k_2 + \dots + k_m$, let us colour numbers $k_1, k_1 + k_2, \dots, k_1 + \dots + k_{m-1}$, $n = k_1 + k_2 + \dots + k_m$ black, and all other positive integers less than n white, see Fig. 10.2 for such a colouring for $n = 4$. Then every representation corresponds to a colouring and vice versa. While n is

Fig. 10.2 Eight ways to represent 4 as a sum of positive integers



always black, there are 2 ways to colour every integer i between 1 and $n - 1$, hence there are 2^{n-1} colourings in total.

However, it is more natural to postulate that the order of summands does not matter, and count representations $3 = 2 + 1$ and $3 = 1 + 2$ as the same one. Representations of n as a sum of other positive integers when the order of summands does not matter are also called *partitions*. Then there are 3 partitions of 3, namely $3 = 3$, $3 = 2 + 1$ and $3 = 1 + 1 + 1$. For $n = 4$, we have

$$4 = 4 = 3 + 1 = 2 + 2 = 2 + 1 + 1 = 1 + 1 + 1 + 1,$$

hence there are 5 partitions of 4. In general, let $p(n)$ be the number of partitions of n . We have just proved that $p(2) = 2$, $p(3) = 3$, $p(4) = 5$. Continuing this way, we can count that $p(5) = 7$, $p(6) = 11$, $p(7) = 15$, $p(8) = 22$, $p(9) = 30$, and so on. However, there is no known simple general formula for $p(n)$, and studying the properties of this function is a non-trivial and interesting problem.

Ramanujan's Theorems

One may observe that $p(4) = 5$ and $p(9) = 30$ are divisible by 5. If one continues the calculation, one may observe that $p(14) = 135$, $p(19) = 490$, $p(24) = 1575$, $p(29) = 4565$ are divisible by 5 as well. This is not a coincidence. About 100 years ago, Ramanujan proved that $p(5n + 4)$ is divisible by 5 for every integer n , see [12]. There is a convenient mathematical notation for this: we write $a \equiv b \pmod{c}$ whenever $b - a$ is divisible by c . In this notation, Ramanujan's theorem states that

$$p(5n + 4) \equiv 0 \pmod{5} \quad n = 0, 1, 2, \dots$$

In fact, Ramanujan also proved that

$$p(7n + 5) \equiv 0 \pmod{7} \quad n = 0, 1, 2, \dots$$

and

$$p(11n + 6) \equiv 0 \pmod{11} \quad n = 0, 1, 2, \dots$$

The next prime after 5, 7, 11 is 13. For about 40 years, it was unknown if a similar statement can be proved for 13, but in the 1960s, Atkin and O'Brien [21] proved that

$$p(17303n + 237) \equiv 0 \pmod{13} \quad n = 0, 1, 2, \dots$$

The Theorem of Ken Ono

The next prime after 13 is 17. Instead of considering the primes one by one, Ken Ono [293] proved in 2000 that for any prime $q \geq 5$ there exist positive integers A and B such that

$$p(An + B) \equiv 0 \pmod{q} \quad n = 0, 1, 2, \dots \quad (10.1)$$

Sequences of the form $An + B, n = 0, 1, 2, \dots$ are called arithmetic progressions. One may ask if $a_n = 5n + 4, n = 0, 1, 2, \dots$ is the only arithmetic progression such that $p(a_n)$ is divisible by 5 for all n . If asked in this form, the answer is clearly “no”, because, for example, the arithmetic progression $b_n = 10n + 4, n = 0, 1, 2, \dots$ is a subsequence of (a_n) , hence, if $p(a_n)$ is divisible by 5 for all n , then trivially the same is true for $p(b_n)$. A family of arithmetic progressions is called *non-nested* if none of them is a subsequence of the union of the other ones. For example, the arithmetic progressions $a_n = 5n + 4, n = 0, 1, 2, \dots$ and $c_n = 7n + 5, n = 0, 1, 2, \dots$ are non-nested. In fact, the theorem of Ken Ono implies that for any q there are *infinitely many* non-nested arithmetic progressions $a_n = An + B, n = 0, 1, 2, \dots$ such that $p(a_n)$ is divisible by q for all n .

Dyson's Rank

Ramanujan's theorem implies that partitions of $5n + 4$ can be divided into 5 groups with an equal number of partitions in each group. For example, $p(9) = 30$, hence partitions of 9 can be divided into 5 groups with 6 partitions in each group. Of course, this may be done in an arbitrary way, but, in 1944, Dyson [123] attempted to find a nice and natural way to perform such division into 5 groups. For this, he introduced the *rank* of each partition as the difference between the largest summand in it and the number of summands. For example, partition $4 = 2 + 1 + 1$ of $n = 4$ has largest summand 2 and the number of summands is 3, hence its rank is equal to $2 - 3 = -1$. Similarly, the ranks of partitions $4 = 4, 4 = 3 + 1, 4 = 2 + 2$, and $4 = 1 + 1 + 1 + 1$ are 3, 1, 0, and -3 , respectively. Dyson noticed that all these ranks have different remainders when divided by 5.

We can do a similar analysis for partitions of 9. Partition $9 = 9$ has rank $9 - 1 = 8$, partition $9 = 8 + 1$ has rank $8 - 2 = 6$, and so on. One can list all 30 partitions of 9 and calculate their ranks. Dyson noticed that, out of these 30 partitions, exactly 6 have

rank divisible by 5 (namely, these are partitions $9 = 7 + 2$, $9 = 5 + 1 + 1 + 1 + 1$, $9 = 4 + 3 + 1 + 1$, $9 = 4 + 2 + 2 + 1$, $9 = 3 + 3 + 3$, $9 = 2 + 2 + 1 + 1 + 1 + 1$, which have ranks $7 - 2 = 5$, $5 - 5 = 0$, $4 - 4 = 0$, $4 - 4 = 0$, $3 - 3 = 0$, and $2 - 7 = -5$, respectively), exactly 6 have a rank which gives remainder 1 after division by 5, and so on, for every $0 \leq r < 5$, there are exactly 6 partitions whose rank gives remainder r after partition by 5. Dyson conjectured that this is true in general, that is, for every n ,

$$N(0, 5; 5n + 4) = N(1, 5; 5n + 4) = \dots = N(4, 5; 5n + 4) = \frac{p(5n + 4)}{5},$$

where $N(r, q; m)$ denotes the number of partitions of m which have rank which gives remainder r after division by q . This conjecture was proved by Atkin and Swinnerton-Dyer [22] in 1954, and this gives a *natural* way to divide $p(5n + 4)$ partitions of $5n + 4$ into 5 groups. In a similar way, Atkin and Swinnerton-Dyer also used Dyson's rank to divide $p(7n + 5)$ partitions of $7n + 5$ into 7 groups.

Divisibility Properties of Dyson's Rank Partition Function

The function $N(r, q; m)$ is called Dyson's rank partition function. Motivated by Ono's Theorem (10.1), one may ask if any general result like (10.1) holds for $N(r, q; m)$ in place of $p(An + B)$. The following theorem [78] gives a positive answer.

Theorem 10.2 *Let t be a positive odd integer, and let q be a prime such that $6t$ is not divisible by q . If j is a positive integer, then there are infinitely many non-nested arithmetic progressions $An + B$ such that for every $0 \leq r < t$ we have*

$$N(r, t; An + B) \equiv 0 \pmod{q^j} \quad n = 0, 1, 2, \dots$$

Because $\sum_{r=0}^{t-1} N(r, t; An + B) = p(An + B)$, Theorem 10.2 (with $j = 1$) implies (10.1). However, it is a much more general result. With $j > 1$, it implies divisibility by any power of prime q , not just by q . More importantly, it states that we can find arithmetic progressions for which not only the sum $\sum_{r=0}^{t-1} N(r, t; An + B)$ is divisible by q^j , but so is every term in this sum!

Reference

K. Bringmann, K. Ono, Dyson's ranks and Maass forms, *Annals of Mathematics* **171**-1, (2010), 419–449.

10.3 Exceptional Times in Percolation Theory

A Mathematical Model for a Water-Resistant Coat

When you walk through the rain, the water can sometimes pass through your clothes and make you wet. This is because clothes have some microscopic “holes” which particles of water can use to go through. If these holes form a connected “channel” through your coat, the water will use this channel to make you wet, see Fig. 10.3a.

This process of movement of fluids through porous materials is called *percolation*. To understand this process, we need a mathematical model describing it. A possible model is to assume that particles of the porous material are moving along a triangular grid, as shown in Fig. 10.3b. Such a grid, which is also called a *triangular lattice*, is just a plane filled by copies of an equilateral triangle with unit side length. The vertices of the triangles are called the vertices of the lattice. If you prefer a formal definition, you can introduce a coordinate plane, and say that a triangular lattice is the set of points in this plane with coordinates $(k + m/2, \sqrt{3}m/2)$ for some integers k and m .

Next we assume that, at any fixed time t , every vertex of the grid is either “empty” or occupied by a particle of the porous material. In the first case, it is “open” for the water to pass through it, while in the second case it is “closed”, and the water cannot pass. Because the particles are moving in a complex unpredictable way, the best we can do is to use probability theory, and assume that every vertex is open with probability p and closed with probability $1 - p$, where $p \in (0, 1)$. We assume that the water can move from one open vertex to another one using the sides of the triangles in the grid, and can therefore pass through the material if there is a connected “channel”, (also called a *path*) through the lattice which uses open vertices only, see Fig. 10.3b. Because particles are tiny, and there are a huge amount of them, a good model is to assume that the lattice is infinite, and we are in fact looking for an infinite path formed from open vertices. If such a path exists, we say that *percolation occurs*. It is known that, for the triangular lattice model, percolation occurs with probability 1 if $p > 1/2$, and does not occur with probability 1 if $p \leq 1/2$. Hence, $p = 1/2$ is called the *critical probability* for the percolation.

Now, if we want to design a water resistant coat, then the “denser” the coat we make the lower p will be in our model. If $p > 1/2$, the coat will leak. On the other

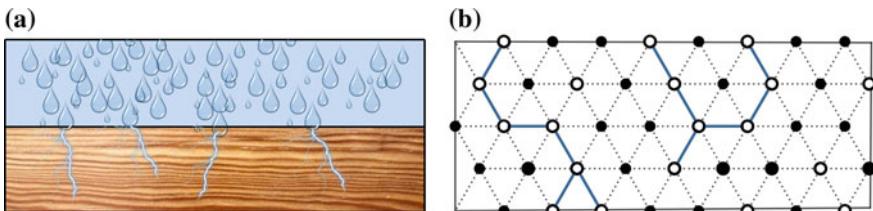


Fig. 10.3 A model for the passage of water through your coat

hand, a coat with $p < 1/2$ would be unnecessarily dense, and therefore unnecessarily expensive. So, it seems that a coat corresponding to the critical probability $p = 1/2$ is the optimal choice. So, let us look at what happens in this “critical” case in more detail.

A Dynamic Model for Critical Site Percolation

The model described above with $p = 1/2$ is called *critical site percolation on the triangular lattice*, and it models what happens to your coat under rain at any fixed time t . As discussed above, in this model percolation does not occur with probability 1, which guarantees that the water has no way to pass through the coat at time t . Does this mean that you really can hope to return to your home dry?

To answer this question, we need a better model, which can explain what happens to your coat dynamically, as time evolves. Of course, the particles are moving, so, in reality, any vertex of the lattice will change its state from “open” to “closed” and back. As a first approximation, we can assume that the particles can move (and hence vertices can change their state) only at some discrete times $t_0, 2t_0, 3t_0, \dots$. At time 0, each vertex is open or closed with equal chance 1/2, and then, at each moment $t_0, 2t_0, 3t_0, \dots$ it can change its state to the opposite one with probability 1/2. In fact, we can measure time in units t_0 , and, in these units, the moments for possible changes are 1, 2, 3, … Then, on average, each vertex will change its state once every 2 time units.

Now, assume that the particles can move (and hence the vertices can change their state) at any time, but the average rate is the same: one change every 2 time units. If we divide each time unit into N parts, we will expect one change per $2N$ parts, hence the probability of change during a time interval of length $\varepsilon = 1/N$ is about $1/(2N) = \varepsilon/2$. Formally, we say that the *rate* of change of a vertex state at time t is the limit, as $\varepsilon \rightarrow 0$, of the probability of change during the time interval $(t, t + \varepsilon)$, divided by ε . We can now formally define our model: we assume that each vertex is open or closed with equal chances at time 0, and then evolves in time in such a way that the rate of change of its state (from open to closed and back) is constant and equal to 1/2 at any time t . All vertices evolve in time independently from each other.

No Chance to Stay Dry

In the model above, if the rain starts at time $t = 0$ and finishes at time $t = 1$, you will become wet if percolation occurs at any time $t \in [0, 1]$. The following theorem [338] states that this will happen with probability 1, so in fact you have no chance to stay dry!

Theorem 10.3 *In the dynamical critical site percolation on the triangular lattice, the set of times $t \in [0, 1]$ at which percolation occurs is nonempty with probability 1.*

The times described in Theorem 10.3 are called *exceptional* times for percolation.

What does this theorem mean for your coat? If you walk through the rain from time 0 to time 1, the static theory guarantees that your chance to become wet at any *specific* time moment t is 0. However, Theorem 10.3 guarantees that you will eventually become wet with probability 1. How can this be? The reason is that there are infinitely many moments of time $t \in [0, 1]$, and, while the chance of becoming wet at any fixed moment t is 0, the *total* chance of being wet is, somewhat informally, the sum of infinitely many zeros, and such a sum can happen to be one. To understand how this may happen, imagine that your computer returns a real number u , chosen uniformly at random from the interval $[0, 1]$, and your computer is programmed to make a sound at time u . Then, for any fixed time $t \in [0, 1]$, the chance of you hearing the sound at exactly time t is 0, but the chance that you will eventually hear the sound is 1.

Generalizations

To prove Theorem 10.3, the authors developed a new theory which provides so-called *quantitative noise sensitivity* estimates for percolations. They use this theory to derive Theorem 10.3, and in fact much more. For example, one may ask what is the chance that the water will find two or more ways to pass through your coat, which are *not connected* to each other. In the described model, this corresponds to the existence of two or more infinite clusters formed from open vertices which are not connected. The authors proved that, with probability 1, this will never happen, even in the dynamic model. In the coat example, this does not mean that the water will find just one way through the coat, so you will be wet in just one place and dry in other places. In fact, it may (and will!) find many channels to go through, and you will be wet everywhere. What the result says is just that all these channels will be connected to each other.

For water-resistant coat designers, Theorem 10.3 tells us that it is a good idea (and not an unnecessary waste of money) to design a dense coat, which corresponds to the percolation model with p strictly less than $1/2$. Of course, this example with coats under rain is somehow artificial and was chosen just for simplicity. The phenomenon of percolation—that is, the movement of fluids through porous materials—occurs everywhere in nature, and is extensively studied in physics, chemistry, and materials science. The corresponding mathematical model is beautiful by itself, and is an active and important area of mathematical research. Theorem 10.3 is an important contribution to this area.

Reference

O. Schramm, J. Steif, Quantitative noise sensitivity and exceptional times for percolation, *Annals of Mathematics* **171**-2, (2010), 619–672.

10.4 Polynomial Parametrization for the Solutions of Diophantine Equations

How to Write Down All the Solutions of an Equation

What does it mean to solve an equation? How to write down the answer? At first, this looks like a trivial question. By “solve an equation” we mean find all its solutions, and prove that there are no more. And, to “write down the answer”, we can just list all the solutions, can’t we? For example, the equation $x^2 = 4$ has two solutions, $x = -2$ and $x = 2$. Similarly, the solution of an equation in two variables, such as $x_1^2 + x_2^2 = 5$, is a pair of values x_1 and x_2 satisfying the equation, for example, $x_1 = 2, x_2 = 1$. In fact, this equation has three more solutions: $x_1 = -2, x_2 = 1$, and $x_1 = 2, x_2 = -1$, and $x_1 = -2, x_2 = -1$, and it is easy to prove that there are no more solutions such that x_1 and x_2 are integers.

In the cases above, we can indeed “write down the answer” by just listing all the solutions. The problem, however, is that some equations have infinitely many solutions, and it is impossible to list them all. Let us try, for example, to find all integer solutions of the equation $3x_1 + 2 = 2x_2 + 1$. Because $3x_1 = 2x_2 - 1$ is odd, x_1 is odd as well, and we can write $x_1 = 2y + 1$ for some integer y . Then $3(2y + 1) + 2 = 2x_2 + 1$ implies that $x_2 = 3y + 1$. With $y = 0$, this leads to a solution $x_1 = 1, x_2 = 1$. With $y = 1$, to a solution $x_1 = 3, x_2 = 4$. With $y = -1$, to a solution $x_1 = -1, x_2 = -2$. In fact, there are infinitely many solutions, one for each integer y , and we cannot explicitly list them all. The best we can do is to just write the whole *family* of solutions in the form

$$x_1 = 2y + 1, \quad x_2 = 3y + 1, \quad y \in \mathbb{Z}, \quad (10.2)$$

where \mathbb{Z} denotes the set of all integers. Formula (10.2) fully describes the infinite family of solutions of our equation, and there are no more integer solutions, hence (10.2) looks like a correct way to write down “the answer” to such an equation. In (10.2), y is called a *parameter*.

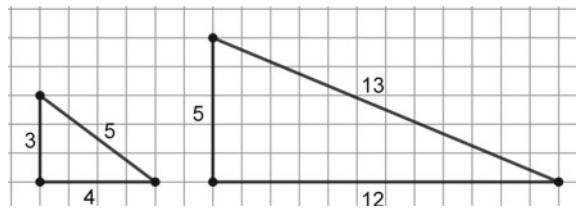
Writing Down Solutions Using Several Parameters

For some equations, it is natural to write the answer using more than one parameter. For example, the famous Pythagorean theorem states that the side lengths x_1, x_2, x_3 of a right triangle satisfy the equation

$$x_1^2 + x_2^2 = x_3^2. \quad (10.3)$$

There exist some right triangles with integer side lengths, for example, with $x_1 = 3, x_2 = 4, x_3 = 5$, or with $x_1 = 5, x_2 = 12, x_3 = 13$, see Fig. 10.4.

Fig. 10.4 Some right triangles with integer side lengths



It is interesting to find all such triangles, but it turns out that there are infinitely many of them, and we need to use formulas with parameters to list them all. It is easy to check that for all integers y_1, y_2, y_3 we have

$$(2y_1y_2y_3)^2 + (y_1^2y_3 - y_2^2y_3)^2 = (y_1^2y_3 + y_2^2y_3)^2,$$

hence

$$x_1 = 2y_1y_2y_3, \quad x_2 = y_1^2y_3 - y_2^2y_3, \quad x_3 = y_1^2y_3 + y_2^2y_3, \quad (y_1, y_2, y_3) \in \mathbb{Z}^3, \quad (10.4)$$

where \mathbb{Z}^3 denotes the set of all triples of integers, is an infinite family of solutions of (10.3) with three parameters. Of course, we can change the order of x_1 and x_2 , and write another family of solutions

$$x_1 = y_1^2y_3 - y_2^2y_3, \quad x_2 = 2y_1y_2y_3, \quad x_3 = y_1^2y_3 + y_2^2y_3, \quad (y_1, y_2, y_3) \in \mathbb{Z}^3, \quad (10.5)$$

but it is not difficult to prove¹ that there are no more: every integer solution (x_1, x_2, x_3) to (10.3) belongs to either family (10.4) or (10.5). For example, the solution $x_1 = 3, x_2 = 4, x_3 = 5$ is (10.5) with $y_1 = 2, y_2 = y_3 = 1$.

Writing Down the Solutions as (a Finite Union of) Polynomial Families

In general, assume that we would like to find all integer solutions of the general equation

$$f(x_1, x_2, \dots, x_k) = 0 \quad (10.6)$$

in k variables. If there are infinitely many solutions, we cannot list them all one by one. Instead, we can hope to find polynomials $P_1(y_1, \dots, y_N), P_2(y_1, \dots, y_N), \dots, P_k(y_1, \dots, y_N)$ in N variables with integer coefficients, such that

$$f(P_1(y_1, \dots, y_N), P_2(y_1, \dots, y_N), \dots, P_k(y_1, \dots, y_N)) = 0, \quad \forall (y_1, y_2, \dots, y_N) \in \mathbb{Z}^N.$$

This would imply that the formula

¹In fact, the proof is presented in Sect. 2.4.

$$x_1 = P_1(y_1, \dots, y_N), x_2 = P_2(y_1, \dots, y_N), \dots x_k = P_k(y_1, \dots, y_N), \quad (y_1, y_2, \dots, y_N) \in \mathbb{Z}^N \quad (10.7)$$

defines an N -parameter family of solutions of (10.6), which is called a *polynomial family*. If we can prove that *all* integer solutions of (10.6) are representable in the form (10.7), we say that “the set of solutions of (10.6) is a polynomial family (10.7)”, and we are done. For Eq.(10.3), the set of solutions cannot be represented as one polynomial family, but can be written as a union of two polynomial families (10.4) and (10.5). In general, whenever the set of solutions of (10.6) is a union of any finite number of polynomial families, we can list all these families, and, in such a way, write down all the solutions.

A Solution to One Equation, with Applications to Others

However, it may be far from trivial to prove (or disprove) that the set of solutions of a given equation is a polynomial family, even for some very simple equations. For example, an old question, which (in a slightly different form) was asked by Skolem [347] in 1938, is whether the set of integer solutions of the equation

$$x_1x_2 - x_3x_4 = 1 \quad (10.8)$$

is a polynomial family. There was very little progress on this question for 70 years, until it was answered positively in the following theorem of Leonid Vaserstein [387].

Theorem 10.4 *The set of all integer solutions of (10.8) is a polynomial family with 46 parameters.*

The importance of Theorem 10.4 goes far beyond solving one specific Eq. (10.8). In fact, this equation was a key to solving many other equations! By either direct application of Theorem 10.4, or using techniques developed to prove it, Vaserstein solved many other interesting equations. For example, he proved that

- (a) For any $n \geq 2$, the set of all integer solutions of the equation $x_1x_2 + x_3x_4 + \dots + x_{2n-1}x_{2n} = 1$ is a polynomial family.
- (b) For any integer D , the set of all integer solutions of the equation $x_1x_2 = x_3^2 + D$ is a finite union of polynomial families.
- (c) For any integer D , the set of all integer solutions of the equation $x_1x_2 + x_3x_4 = D$ is a finite union of polynomial families.
- (d) The set of all integer solutions of the equation $x_1^2 + x_2^2 = x_3^2 + 3$ is a union of two polynomial families.

Representation of Primitive Solutions

If a solution (x_1, x_2, x_3) of Eq. (10.3) has a common factor $d \geq 2$, then $(x_1/d, x_2/d, x_3/d)$ is also a solution of (10.3). In other words, the common factor d can be “can-

celled out". For this reason, solutions with a common factor are somewhat "uninteresting". For example, the solution $x_1 = 300, x_2 = 400, x_3 = 500$ is just the solution $x_1 = 3, x_2 = 4, x_3 = 5$, multiplied by common factor 100. One may sometimes wish to exclude such solutions from consideration, and list only *primitive* solutions of an equation. A solution (x_1, \dots, x_k) of (10.6) is called primitive if x_1, \dots, x_k have no common factor greater than 1. Can we find a polynomial family (10.7) to list all the *primitive* solution of a given equation? Theorem 10.4 can be used to answer this question for many interesting equations, for example:

- (e) The set of all primitive solutions of any linear system of equations with integer coefficients either consists of two solutions or is a polynomial family.
- (f) The set of primitive solutions of Eq.(10.3) is the union of four polynomial families with 95 parameters each.
- (g) For any integer D , the set of all primitive solutions of the equation $x_1x_2 + x_3x_4 = D$ is a polynomial family with 92 parameters.
- (h) For any integer D , the set of all primitive solutions of the equation $x_1x_2 = x_3^2 + D$ is a finite union of polynomial families with 46 parameters each.
- (i) For arbitrary integers a, b, c and any integers α, β, γ the set of primitive solutions to the equation $ax_1^\alpha + bx_2^\beta = cx_3^\gamma$ can be covered by a finite (possibly, empty) set of polynomial families.

Negative Results and Open Questions

On the other hand, results like the above are not possible for every equation. For example, for any square-free integer D (D is called square-free if it cannot be represented in the form $D = ab^2$ for some integers a and $b \geq 2$), the set of all integer solutions of the equation $x_1^2 - Dx_2^2 = 1$ is *not* a union of any finite number of polynomial families. For some other interesting equations, like, for example, $x_1^3 + x_2^3 + x_3^3 = 1$, or $x_1^3 + x_2^3 + x_3^3 + x_4^3 = 0$, the question of whether their solution sets are a finite unions of polynomial families is still an open problem.

Reference

L. Vaserstein, Polynomial parametrization for the solutions of Diophantine equations and arithmetic groups, *Annals of Mathematics* **171**-2, (2010), 979–1009.

10.5 The Order of Growth of Variance in the Asymmetric Simple Exclusion Process

Understanding and Modelling Traffic Flow in a City

When the government of a city is planning to close some roads for repair, they need to understand whether the remaining roads will be sufficient for normal traffic flow, with

no significant traffic jams. If the cars are moving smoothly with constant velocity, then even a single one-way single-lane road can be used to pass a lot of cars every minute. The problem, however, is that the cars may slow down or stop for various reasons, e.g. because of some technical issues, to let passengers in or out, to avoid pedestrians, etc. In a single-lane road, even one stopping car forces many cars after it to stop as well. Taking this into account, how can we accurately estimate the number of cars which can pass through such a road, say, every hour?

A Model for Car Movement and Interaction: TASEP

To answer the question above, we need to develop a mathematical model to describe this phenomenon. We can model our single-lane road as a coordinate line, and assume that each car can occupy only positions with integer coordinates $\dots - 2, -1, 0, 1, 2, \dots$. We first define an initial configuration at time $t = 0$. To do this, let us fix some real number $\sigma \in (0, 1)$, and imagine a coin with two sides “Yes” and “No”, such that, when we toss this coin, it always shows “Yes” with probability σ and “No” with probability $1 - \sigma$, and this result is independent in each new experiment. For every position $i \in \mathbb{Z}$ (\mathbb{Z} denotes the set of all integers), let us toss this coin, and put a car at position i if the coin showed “Yes”, and leave the position i empty otherwise. As a result, on every long interval of length L , we expect about σL cars, so that σ is interpreted as the “density” of cars on the road.

Next we need to define the law for movement of the cars. We assume that a car at any position $i \in \mathbb{Z}$ can either stay there or move to a position $i + 1$, provided that there are no other cars at that position. To model a movement at “random” times, we assume that each car has a private clock, ringing from time to time, such that the chance that it rings during any time interval $(t, t + \varepsilon)$ is the same for all t . All clocks ring independently of each other. If a clock rings at any time t , the corresponding car moves from its current position i to the position $i + 1$, if this position is free, and otherwise stays at i . The model we have just described is called the *totally asymmetric simple exclusion process* (TASEP).

A Model for Particle Movement and Interaction: ASEP

TASEP is a special case of the following more general model for particle movement and interaction from statistical physics. In this model, the initial configuration of particles is the same as described above (that is, particles occupy positions with integer coordinates with some density σ), but the particles, unlike cars, can move in *both* directions: from every position $i \in \mathbb{Z}$, a particle can move to either position $i - 1$ (to the left) or to $i + 1$ (to the right). Fix some real number $p \in (1/2, 1]$, and let $q = 1 - p$ (note that q is less than p). Imagine that each particle has two private clocks, named L and R , and both ring from time to time. For the clock L , the chance that it rings during any time interval $(t, t + \varepsilon)$ is the same for all t , and is equal to

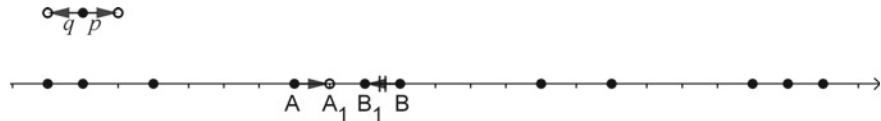


Fig. 10.5 A model for particle movement and interaction

$q\varepsilon + o(\varepsilon)$, where $o(\varepsilon)$ is a function such that $\lim_{\varepsilon \rightarrow 0} \frac{o(\varepsilon)}{\varepsilon} = 0$. In other words, clock L rings with “constant rate” q rings per unit of time. Clock R is similar but rings with rate p rings per unit of time. All clocks ring independently of each other. If clock L rings at any time t , the corresponding particle moves from its current position i to the left, to position $i - 1$, if this position is free (that is, there are no particles currently at $i - 1$), and otherwise stays at i . Similarly, if clock R rings, a particle at position i moves to the right, to position $i + 1$, if this position is free, and stays at i otherwise. Figure 10.5 demonstrates the situation in which clock R rings for a particle at position A , so that it moves to position A_1 . After some time, clock L rings for a particle at position B , but, because position B_1 on the left is occupied, no movement happened.

This model for particle movement and interaction is called the *asymmetric simple exclusion process* (ASEP). It is fully described by parameters σ , p , and $q = 1 - p$. While parameter σ represents the density of particles on the line, parameters p and q describe how often particles will move to the right and left, respectively. The TASEP model for car movement is the special case of ASEP with $p = 1$ and $q = 0$.

How Many Particles Will an Observer Meet?

Now assume that there is an observer which, starting from position 0 at time 0, is travelling along the line at constant speed v during the time interval $(0, t]$, with the convention that $v > 0$ corresponds to movement from left to right, and $v < 0$ if the movement is from right to left. Then the observer will be at coordinate vt at time t . Let us denote by $[vt]$ the last integer point the observer has passed. In other words, $[vt]$ is the largest integer in $[0, vt]$ if $vt \geq 0$, and the smallest integer in $[vt, 0]$ if $vt \leq 0$. Next, let $J_+^v(t)$ be the number of particles which have passed the observer from left to right, that is, particles which started at position $i \leq 0$ at time 0 but finished in position $j \geq [vt] + 1$ at time t . Similarly, let $J_-^v(t)$ be the number of particles which have passed the observer in the opposite direction, that is, started at position $i \geq 1$ at time 0 but finished at $j \leq [vt]$ at time t . Finally, let $J^v(t) = J_+^v(t) - J_-^v(t)$ be the “net” number of particles which have passed the observer from left to right. Our goal is to understand how large $J^v(t)$ is in the described model.

Estimate of the “Average”, or Expected Value

Because our model involves random events, $J^v(t)$ is a random variable, that is, it may be larger or smaller just by chance if we repeat the experiment several times. Hence, we cannot compute/predict its exact value. However, we can compute/estimate the *average* value of $J^v(t)$ we would get if we repeat the experiment several times. For example, if we toss a standard die, the result is a random variable X taking values 1, 2, 3, 4, 5, 6 with equal chances. If we repeat such an experiment N times, we can expect that every outcome will occur about $N/6$ times, hence the average outcome of all experiments will be about $\frac{1}{N}(1 \cdot N/6 + 2 \cdot N/6 + 3 \cdot N/6 + 4 \cdot N/6 + 5 \cdot N/6 + 6 \cdot N/6) = 3.5$. This value 3.5 is called the *expectation* of the random variable X and is usually denoted by $E[X]$. More generally, for any discrete random variable X taking values x_1, x_2, \dots, x_n with probabilities p_1, p_2, \dots, p_n , respectively, the expectation is $E[X] = \sum_{i=1}^n p_i x_i$. For an arbitrary random variable X , the expectation is defined as $E[X] = \int_{-\infty}^{\infty} x dF(x)$, where, for any real number x , $F(x)$ denotes the probability that $X \leq x$.

In our case, calculations show that

$$E[J^v(t)] = t(p - q)\sigma(1 - \sigma) - \sigma[vt]. \quad (10.9)$$

If we program a computer simulation of our model, run it many times, and compute $J^v(t)$ every time, that value $E[J^v(t)]$ is what we would get *on average*.

How close is $E[J^v(t)]$ in (10.9) to the *actual* value of $J^v(t)$?

Formula (10.9) can be treated as a *prediction* of the actual number $J^v(t)$ of particles the observer will meet, and this prediction works well on average. But how accurate is it? For example, in the die experiment described above, we can get the result $X = 1$ of the die toss, which is significantly lower than $E[X] = 3.5$, as well as the result $X = 6$, which is significantly higher than $E[X]$. On the other hand, imagine a fair coin with the numbers 3.499 and 3.501 written on each of its sides. If we toss such a coin, the result would be a random variable Y with $E[Y] = 3.5$, but, in this case, the actual values we could get in the experiment (3.499 or 3.501) are pretty close to the “predicted” value $E[Y]$.

To understand how close the actual values of a random variable X are to its average/predicted value $E[X]$, we can take the difference $X - E[X]$, and calculate the average of the square of this difference, that is, $E[(X - E[X])^2]$. This quantity is known as the *variance* of X , denoted $\text{Var}(X)$, and is a fundamental quantity associated with any random variable. If the variance of X is small, we may be sure that $E[X]$ is likely to be an accurate prediction for the actual value of X we get in the experiment. For example, if $\text{Var}(J^v(t))$ is “small”, $E[J^v(t)]$ given by (10.9) is an accurate prediction of the *actual* value of $J^v(t)$. But is $\text{Var}(J^v(t))$ *really* small? Can we estimate it?

Estimating the Variance of $J^v(t)$

In 1994, Ferrari and Fontes [151] proved that

$$\lim_{t \rightarrow \infty} \frac{\text{Var}(J^v(t))}{t} = \sigma(1 - \sigma)|(p - q)(1 - 2\sigma) - v|, \quad (10.10)$$

where $|\cdot|$ denotes the absolute value. If $v \neq (p - q)(1 - 2\sigma)$, then $C = \sigma(1 - \sigma) \cdot |(p - q)(1 - 2\sigma) - v|$ is some positive constant, and (10.10) implies that $\text{Var}(J^v(t)) \approx Ct$ for large t , that is, the variance in question grows linearly with time. However, (10.10) also implies that $\lim_{t \rightarrow \infty} \frac{\text{Var}(J^v(t))}{t} = 0$ for the special speed $v = v^*$, where

$$v^* = (p - q)(1 - 2\sigma).$$

That is, at speed $v = v^*$ the variance of $J^v(t)$ grows slower than linearly in t , hence, in this case, $E[J^v(t)]$ gives a more accurate prediction of $J^v(t)$ than it gives for $v \neq v^*$.

But how close is $J^v(t)$ to $E[J^v(t)]$ at the critical speed $v = v^*$? How fast does $\text{Var}(J^v(t))$ grow in this case? Experiments and informal physical reasoning predicted that in this case it should grow as $t^{2/3}$. In 2010, Balázs and Seppäläinen [33] gave a formal mathematical proof of this informal prediction.

Theorem 10.5 *In the model described above, for $v = v^*$, there exist constants $t_0 > 0$ and $C < \infty$ such that, for all $t \geq t_0$,*

$$C^{-1}t^{2/3} \leq \text{Var}(J^v(t)) \leq Ct^{2/3}.$$

Reference

M. Balázs, T. Seppäläinen, Order of current variance and diffusivity in the asymmetric simple exclusion process, *Annals of Mathematics* **171**-2, (2010), 1237–1265.

10.6 Divergent Square Averages in Ergodic Dynamical Systems

Mathematical Models for Systems Which Evolve in Time

In essentially all areas of science we need to understand how “something evolves with time”. For example, in physics it is interesting to understand how planets and stars are moving in space; in biology, one studies how living organisms evolve with time; in economics, we can study the evolution of prices of various goods, etc. Can we build a mathematical theory which is general enough to study all such phenomena?

To start building such a theory, let us introduce the set X of “all possible states of the world”, that is, all possible “ways” our system may in principle look like. For example, when we study planetary motion, then the “state of the world” is the

position and velocity of every planet, every star, and, more generally, every object in space. In economics, X may be the set of all possible prices of all goods in some market, etc.

At any fixed time t , our system is at some “state of the world” $x(t) \in X$. In particular, let x_0, x_1, x_2, \dots be the sequence of states of the world at some discrete time moments $t = 0, 1, 2, 3, \dots$. The crucial assumption we make is that the previous state of the world fully determines the next one, that is, there is a function $T : X \rightarrow X$ such that $x_{n+1} = T(x_n)$, $n = 0, 1, 2, \dots$. Intuitively, T represents the laws of nature according to which the system evolves. For example, if we know the position and velocity of every object in space at time $t = n$, we can (in principle) use Newton’s laws of motion to compute such positions and velocities at time $t = n + 1$. The fact that T does not depend on n represents the intuition that the laws of nature are the same at all times.

If we know the initial “state of the world” $x_0 \in X$, and the “law of evolution” T , we can compute $x_1 = T(x_0)$, $x_2 = T(x_1) = T(T(x_0))$, $x_3 = T(T(T(x_0)))$, and so on. In general, we denote by $T^n(x)$ the expression $T(T(\dots T(x) \dots))$, where T is iterated n times. Then $x_n = T^n(x_0)$, $n = 0, 1, 2, \dots$.

Estimating Average Values with Respect to Time

Now assume that we need to compute some specific quantity of interest, expressed as a real number, for example, the distance between Earth and the Moon. If we know the state of the world $x \in X$, we can obviously do this (because the state of the world determines everything), hence our “quantity of interest” is just a function $f : X \rightarrow \mathbb{R}$, where \mathbb{R} is the set of all real numbers. To study how our “quantity of interest” evolves with time, we need to study a sequence of real numbers,

$$f(x_0), f(T(x_0)), f(T^2(x_0)), \dots, f(T^n(x_0)), \dots \quad (10.11)$$

If you type “the distance between Earth and the Moon” into Google search, it returns the value “384,400 km”. Of course, everyone understands that in fact this distance is not a constant, and is subject to some time fluctuations: sometimes the Moon is closer to Earth, and sometimes further away. What Google returns to you is the (approximate) *average* distance we would get if we observed the Moon over some long period of time. In general, we are often interested not in the whole sequence (10.11), describing how our quantity of interest depends on time, but only in the *average* value of f after N observations

$$\bar{f}(x_0, N) := \frac{1}{N} \sum_{n=1}^N f(T^n(x_0)) \quad (10.12)$$

If we measure the distance to the Moon $N = 3$ times on January 1st, 2nd, and 3rd, and take the average out of the three results we get, it may be not a very accurate

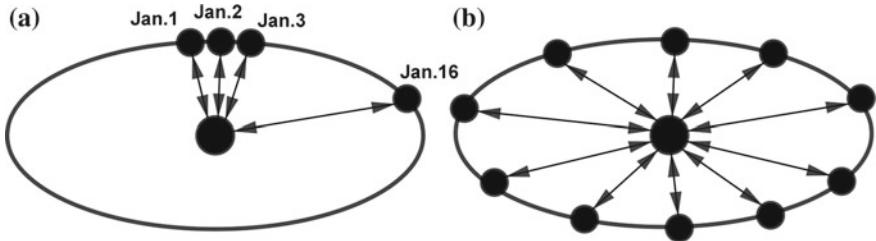


Fig. 10.6 Estimating the average distance to the Moon

estimate, because the distance may be very different in two weeks, see Fig. 10.6a. Intuitively, the value $\bar{f}(x_0, N)$ becomes a better reflection of reality only when N becomes large, the larger the better, see Fig. 10.6b. In fact, it is natural to study the limit

$$\lim_{N \rightarrow \infty} \bar{f}(x_0, N) \quad (10.13)$$

when the number N of observations goes to infinity. In the case of the Earth-Moon distance, this limit does indeed exist and is approximately equal to 384,400 km. In general, however, we (theoretically) may be faced with the situation where, for example, for some function f , $\bar{f}(x_0, N) = 1$ for odd N , and $\bar{f}(x_0, N) = 2$ for even N , hence the limit does not exist. A famous and important theorem called Birkhoff's Ergodic Theorem [57] states that, under some conditions on f and T , this limit is guaranteed to exist for “almost all” initial configurations $x_0 \in X$.

Measure Spaces and the Concept of “Almost All”

To understand what is meant by “some conditions on f and T ” and by “almost all initial configurations”, we need some more definitions. First, we define the words “almost all”. We start with a famous example, that “almost all real numbers on $[0, 1]$ are irrational”. This is because it is known that all rational numbers can be enumerated, that is, there is a sequence $x_1, x_2, \dots, x_n, \dots$ containing *all* rational numbers. We can then select $\varepsilon > 0$ and cover each x_n by a small interval of length $\varepsilon/2^n$, so that the total length of all such intervals is ε . Subsets of $[0, 1]$ which can be covered by intervals with total length ε for any $\varepsilon > 0$ are called “sets of Lebesgue measure 0”, and, if some property holds for all $x \in [0, 1]$ except possibly for a set of measure 0, we say that it holds for “almost all” x .

Let us return to our general setting. Let X be an abstract set of “all possible states of the world”, and let S be any subset of X (for example, the set of all initial configurations $x_0 \in X$ for which Birkhoff's Ergodic Theorem holds). We want to say that “ $x \in S$ for almost all $x \in X$ ”. To be able to say anything like this, we need to introduce some way to measure “how big” subsets of X are, analogous to the notion of “length” on $[0, 1]$. As a first step, we divide all subsets $A \subset X$ into two classes:

“measurable” and “not measurable”, in such a way that (i) X is measurable, (ii) if A is measurable, then so is $X \setminus A$ ($X \setminus A$ is the set of all $x \in X$ which do not belong to A) and (iii) if A_1, A_2, A_3, \dots are all measurable, then so is $A = A_1 \cup A_2 \cup A_3 \cup \dots$. Then, to every measurable set $A \subset X$, we associate a real number $\mu(A)$, called its *measure*, in such a way that (i) $\mu(A) \geq 0$ for all measurable A , (ii) $\mu(\emptyset) = 0$, where \emptyset denotes the empty set, and (iii) $\mu(A) = \sum_{n=1}^{\infty} \mu(A_n)$, whenever A_1, A_2, A_3, \dots are pairwise disjoint measurable sets and $A = A_1 \cup A_2 \cup A_3 \cup \dots$. A set X equipped with a measure μ as defined above is called a *measure space*. In addition, we assume that $\mu(X) = 1$. Now, we say that some property (for example, Birkhoff’s Ergodic Theorem) holds for “almost all” $x_0 \in X$ if the set S of all x_0 for which it holds has measure 1, or, equivalently, the set $X \setminus S$ of all x_0 for which the property does *not* hold has measure 0.

Ergodic Dynamical Systems

We will next define what we mean by “some conditions on f and T ”, and we start with conditions on T . Imagine that there is some set $A \subset X$ of special states of the world such that $T(x) \in A$ whenever $x \in A$. Then, if x_0 belongs to A , then so is $x_1 = T(x_0)$, hence so is $x_2 = T(x_1)$, and so on. In other words, once our system has started in A , it never leaves it. Our main condition on T will aim to exclude such “degenerate” situations.

Formally, if X is a measure space, $T : X \rightarrow X$ is called a measure-preserving transformation if for any measurable $A \subset X$, the set $T^{-1}(A) := \{x \in X \mid T(x) \in A\}$ is measurable and $\mu(T^{-1}(A)) = \mu(A)$. Next, a measure-preserving transformation $T : X \rightarrow X$ is called *ergodic* if, for any measurable set A such that $T^{-1}(A) = A$, we either have $\mu(A) = 0$ or $\mu(A) = 1$. In other words, if $0 < \mu(A) < 1$, then $T^{-1}(A) \neq A$, and the “degenerate behaviour” described above cannot happen. A measure space X , together with an ergodic transformation $T : X \rightarrow X$, is called an *ergodic dynamical system*. For example, if $X = [0, 1]$ with the usual measure, then the transformation $T(x) = x^2$ is not measure-preserving, because, for example, for $A = [0, 0.25]$, we have $T^{-1}(A) = \{x \in [0, 1] \mid x^2 \in [0, 0.25]\} = [0, 0.5]$, hence $\mu(T^{-1}(A)) = 0.5 \neq 0.25 = \mu(A)$. Next, the transformation $T(x) = 1 - x$ is a measure-preserving transformation, but not ergodic, because $T^{-1}(A) = A$ for $A = [0.4, 0.6]$ of measure $0 < \mu(A) = 0.2 < 1$. On the other hand, $X = [0, 1]$ with $T(x) = (x + \sqrt{2}) - [x + \sqrt{2}]$, where $[z]$ denotes the largest integer not exceeding z , defines an ergodic dynamical system.

Absolutely Integrable Functions on Measure Spaces

Finally, we formulate the condition on f . Informally, it states that f should not “grow too fast”. Formally, we say that a function $g : X \rightarrow \mathbb{R}$ is *simple* if there exists a sequence of pairwise disjoint measurable sets A_1, A_2, \dots such that

$X = A_1 \cup A_2 \cup \dots$ and a sequence of real numbers a_1, a_2, \dots such that $g(x) = a_n$, $\forall x \in A_n$, $n = 1, 2, 3, \dots$. A simple function g with non-negative values is called *integrable* if $I(g) := \sum_{n=1}^{\infty} |a_n| \mu(A_n) < \infty$. A function f is called *measurable* if $f^{-1}(A)$ is a measurable set whenever A is. A measurable function f is called *absolutely integrable* if there exists a real number $B < \infty$ such that $I(g) \leq B$ for every simple integrable function g such that $0 \leq g(x) \leq |f(x)|$, $\forall x \in X$. The set of all absolutely integrable functions f is denoted L^1 . For example, for the measure space $X = [0, 1]$, the function $f(x) = x^2$ is absolutely integrable, and $B = \int_0^1 x^2 dx = \frac{1}{3}$. On the other hand, the function $f(x) = 1/x$, $x \in (0, 1]$ (we can assign $f(0) = 0$ to make it well-defined on $[0, 1]$) grows too fast as $x \rightarrow 0$, and it is not absolutely integrable.

We can now formulate Birkhoff's Ergodic Theorem: if X is a measure space, $T : X \rightarrow X$ is ergodic, and $f : X \rightarrow \mathbb{R}$ is absolutely integrable, then the limit (10.13) exists² for almost all $x_0 \in X$.

Divergent Square Averages

In (10.12), we used the values of f at all times $t = 1, 2, \dots, N$ to calculate the average. However, what if we perform measurements only at some specific moments of time $0 < n_1 < n_2 < \dots$? For example, consider the sequence $0 < 1 < 4 < 9 < 16 < 25 < \dots$ of perfect squares, and introduce the average of f at the corresponding times:

$$h_f(x_0, N) := \frac{1}{N} \sum_{n=1}^N f(T^{n^2}(x_0)).$$

Can we prove that the limit $\lim_{N \rightarrow \infty} h_f(x_0, N)$ exists for almost all x_0 in this case? This question was open for almost 30 years, until it was answered negatively by the following theorem [80].

Theorem 10.6 *For every ergodic dynamical system, there exists an absolutely integrable function f such that the limit $\lim_{N \rightarrow \infty} h_f(x_0, N)$ fails to exist for all $x_0 \in A$, where A is a set of positive measure.*

Reference

Z. Buczolich and D. Mauldin, Divergent square averages, *Annals of Mathematics* **171**-3, (2010), 1479–1530.

²In fact, the theorem also states that this limit does not depend on x_0 .

Fig. 10.7 Primes with even and odd sums of digits

2	3	5	7	11	13	17	19	23	29	31	37
41	43	47	53	59	61	67	71	73	79	83	89
97	101	103	107	109	113	127	131	137	139	149	151
157	163	167	173	179	181	191	193	197	199	211	223
227	229	233	239	241	251	257	263	269	271	277	281

10.7 Divisibility Properties of Sums of Digits of Prime Numbers

Mysteries Hidden in the Sequence of Prime Number

One of the most interesting challenges in number theory, and perhaps in the whole of mathematics, is to understand the mysteries around the prime numbers, positive integers which cannot be non-trivially factorized. The sequence of prime numbers starts with

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37 \dots$$

and contains infinitely many terms. Some of the very basic questions about this sequence remain unanswered. For example, we do not know if there are infinitely many integers p such that both p and $p + 2$ are primes, and this is one of the most famous open questions in the whole of mathematics.

Let us consider another question about prime numbers, which is also easy to understand. Let us colour a prime number p black if the sum of digits in its decimal expansion is even, and white if this sum is odd. For example, for $p = 11$, the sum of digits is $1 + 1 = 2$, hence 11 should be coloured black. In contrast, for $p = 23$, the sum of digits is $2 + 3 = 5$, hence 23 will be coloured white. The colouring of the first 60 primes is presented in Fig. 10.7. Now, the question is: are there more black primes or white primes?

How to Compare the Number of Terms in Infinite Sequences

We need some care to formulate this question rigorously. After all, it is not difficult to guess (but more difficult to prove) that there are infinitely many black primes, and also infinitely many white ones. So, can we in principle say that one infinite sequence of integers contains “more” terms than another one?

To better understand this question, let us consider the following colouring of *all* positive integers (not necessarily primes): let us colour positive integers divisible by 10 in red, and all other positive integers in blue. Then, obviously, there are infinitely many integers coloured in each of the colours. However, intuitively, there are much more blue integers than red ones.

To formalize this intuition, let us look at the first n positive integers for some large but finite n . For simplicity, assume that n is divisible by 10. Then, among the first n

positive integers, there are only $n/10$ red ones, and $9n/10$ blue ones. In other words, 10% of all positive integers up to n are red, and 90% are blue. For n not divisible by 10, the situation is similar, up to some very small error. Hence, we can conclude that the (infinite) set of blue integers is “larger” than the (infinite) set of red ones, in fact 9 times “larger”.

So, More Black Primes or More White Ones?

Returning to our question about black and white primes, we can proceed similarly. For any integer n , let $\pi(n)$ be the number of prime numbers not exceeding n , and let $\pi_b(n)$ and $\pi_w(n)$ be the number of black and white primes not exceeding n , respectively. Then $\pi_b(n) + \pi_w(n) = \pi(n)$, and, because $\pi_b(n)$ and $\pi_w(n)$ are now finite numbers, we can legitimately compare them and decide which one is larger.

So, what would you expect, more black primes or more white ones? In fact, because half the integers are even and half are odd, it is intuitively “clear” that the sum of digits in a randomly chosen prime number has an “equal chance” to be even and odd, so we would expect that about half of all primes should be coloured black, and about half coloured white. In other words, we expect that

$$\pi_b(n) \approx \pi_w(n) \approx \frac{\pi(n)}{2}.$$

How Many Primes Have Sum of Digits Divisible by 3?

As discussed above, we expect that about half of all primes have an even sum of digits. One may ask, is it also true that about one third of all primes have sum of digits divisible by 3, one fourth of all primes have sum of digits divisible by 4, and so on? If you think about this for a minute, you will quickly find out that this is *not* always true. For example, the set of primes with sum of digits divisible by 3 is *not* a third of all primes. In fact, there are no primes with this property except for $p = 3$. This is because the famous divisibility rule states that the sum of digits of any number k is divisible by 3 if and only if the number itself is divisible by 3. But no prime $p > 3$ is divisible by 3 by definition, hence the sum of digits of p is not divisible by 3.

Non-decimal Number Systems

So far, we have discussed the sum of digits of prime numbers in our usual decimal system. In it, integers 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 are denoted using a new symbol (digit) for each integer, and, for the next integer, ten, we do not introduce a new symbol, but write “10” instead. Historically, people started to use the decimal system because they have ten fingers on two hands. However, they might well use the fingers of one

hand only, use only five digits 0, 1, 2, 3, 4, and denote by “10” the next integer, five. Then six would be written as 11, seven as 12, eight as 13, nine as 14, ten as 20. We can then continue 21, 22,, and, when we reach 44, the next integer would be written as 100, and so on. In this system, the first primes are written as

$$2, 3, 10, 12, 21, 23, 32, 34, 43, 104, 111, 122 \dots$$

You can check that, in this case, the sum of digits of all $p > 2$ is always odd. This is because in this number system the sum of digits is even if and only if the number itself is even, and there are no even primes other than 2. On the other hand, the divisibility by 3 rule does not work here. For example, the number seven is written as 12, and has sum of digits $1 + 2 = 3$, but seven itself is not divisible by 3. In fact, seven is prime. In general, we now have many primes (3, 12, 21, 111, ...) with sum of digits divisible by 3. Moreover, we also have some primes (10, 34, 43, ...) with sum of digits which gives remainder 1 when divided by 3, and some primes (2, 23, 32, 104, 122, ...) with sum of digits which gives remainder 2. Are all primes approximately equally distributed among these three groups?

A Problem of Gelfond

In the decimal system, we may represent the number 2745 as $2 \cdot 10^3 + 7 \cdot 10^2 + 4 \cdot 10 + 5$. More generally, any number n written as $a_1 a_2 \dots a_k$ (here a_1, a_2, \dots, a_k are digits used to write down n) is equal to

$$n = a_1 \cdot 10^{k-1} + a_2 \cdot 10^{k-2} + \dots + a_{k-1} \cdot 10 + a_k.$$

Similarly, for any integer $q \geq 2$, any integer n can be represented as

$$n = b_1 \cdot q^{k-1} + b_2 \cdot q^{k-2} + \dots + b_{k-1} \cdot q + b_k$$

where k is some positive integer and b_1, \dots, b_k are integers between 0 and $q - 1$. The sum $S_q(n) = b_1 + b_2 + \dots + b_k$ is called *the sum of digits of n in the q-ary system*. In 1968, Gelfond [161] asked to investigate, for every pair of integers q and m , which part of all primes p has sum of digits $S_q(p)$ divisible by m , which part has sum of digits giving remainder 1 after division by m , and so on.

In particular, for which values of q and m does the m -th part of all primes have sum of digits divisible by m ? As we saw above, this is not the case for $q = 10$ and $m = 3$, nor for $q = 5$ and $m = 2$. More generally, whenever $q - 1$ and m has a common divisor $d > 1$, the answer will always be negative. This is because in this case the “divisibility by d ” rule holds: $S_q(p)$ is divisible by d if and only if p is divisible. However, no primes $p > d$ are divisible by d . Numbers with no common divisors greater than 1 are called *relatively prime*. The discussion above motivates the assumption that the numbers $q - 1$ and m are relatively prime.

The Solution

There was essentially no progress on Gelfond's question for more than 40 years, and then it was fully resolved by Mauduit and Rivat [265] in 2010.

Theorem 10.7 *Let $m \geq 2$ and $q \geq 2$ be integers such that $q - 1$ and m are relatively prime, and let a be an integer such that $0 \leq a \leq m - 1$. Let $\pi_{q,m,a}(n)$ be the number of primes p less than n such that the sum of digits of n in the q -ary system gives remainder a after division by m . Then there exists constants $C_{q,m} > 0$ and $\sigma_{q,m} > 0$, which depend on q and m , such that the inequality*

$$\left| \pi_{q,m,a}(n) - \frac{\pi(n)}{m} \right| \leq C_{q,m} n^{1-\sigma_{q,m}}$$

holds for all n .

In particular, case $m = 2$ of Theorem 10.7 implies the existence of constants $C > 0$ and $\sigma > 0$ such that the inequalities

$$|\pi_b(n) - \pi(n)/2| \leq Cn^{1-\sigma}, \quad \text{and} \quad |\pi_w(n) - \pi(n)/2| \leq Cn^{1-\sigma} \quad (10.14)$$

hold for all n , where $\pi_b(n)$ and $\pi_w(n)$ are the numbers of black and white primes up to n , respectively.

Just for illustration, assume for a moment that (10.14) holds with constants $C = 1$ and $\sigma = 1/2$. This would imply that, for $n = 10^{10}$ (ten billion), $|\pi_b(n) - \pi(n)/2| \leq n^{1/2} = 100,000$. Should this really be considered as "small"? Well, the point is that there are $\pi(10^{10}) = 455,052,511$ primes less than ten billion, and the inequality $|\pi_b(n) - \pi(n)/2| \leq 100,000$ implies that, in the very worst case, there may be 227,676,255 black primes and 227,476,256 white ones, or vice versa. As you can see, compared to hundreds of millions of black and white primes, a difference of one or two hundred thousand looks tiny. It implies that the percentage of white primes would be between 49.967% and 50.033%, and the same is true for black ones.

More generally, for any positive constants C and σ (not necessarily for $C = 1$ and $\sigma = 1/2$ as in the example above), and for any $\varepsilon > 0$, however small, the famous prime number theorem guarantees that $\varepsilon\pi(n) > Cn^{1-\sigma}$, if we select n large enough. Then (10.14) implies that $|\pi_w(n) - \pi(n)/2| \leq \varepsilon\pi(n)$, which means that the percentage of white primes is between $50\% - \varepsilon$ and $50\% + \varepsilon$.

In general, for large n , $\pi(n)$ is much larger than $C_{q,m}n^{1-\sigma_{q,m}}$, and Theorem 10.7 implies that, if we colour all primes by m colours depending on what remainder $S_q(n)$ gives after division by m , then the number of primes in each colour class will be approximately the same. So, one interesting mystery about the sequence of prime numbers is now resolved. However, there are many more left.

Reference

C. Mauduit and J. Rivat, Sur un probleme de Gelfond: la somme des chiffres des nombres premiers, *Annals of Mathematics* **171**-3, (2010), 1591–1646.

10.8 Prime Values of Linear Equations

Systems of Linear Equations: Existence of a Solution

One of the easiest mathematical exercises one can imagine is solving a linear equation. For example, you can easily solve the equation $2x + 3 = 7$ and find that it has a unique solution $x = 2$. Not much more difficult is to solve a system of several such equations, such as

$$\begin{cases} 2x + 3y = 7 \\ x - 4y = -2 \end{cases}$$

In this example, we can just express $x = 4y - 2$ from the second equation, and substitute into the first one, to get $2(4y - 2) + 3y = 7$, or $11y = 11$, which implies that $y = 1$ and $x = 4y - 2 = 2$.

Sometimes, systems of linear equations may have no solutions. For example, consider the system

$$\begin{cases} x - y - 2 = 0 \\ -2x + 2y + 5 = 0 \end{cases}$$

From the first equation, $x = y + 2$. Substituting this into the second one gives $-2(y + 2) + 2y + 5 = 0$, which simplifies to $1 = 0$, a contradiction. Why did this happen? Let us write $\psi_1(x, y) = x - y - 2$ and $\psi_2(x, y) = -2x + 2y + 5$ for the left-hand sides of the first and the second equations, respectively. Then the problem arose because ψ_1 and ψ_2 happened to be related in a linear way: $\psi_2(x, y) = -2\psi_1(x, y) + 1$. This implies that they cannot be both zero, because this would lead to $0 = -2 \cdot 0 + 1$, a contradiction. In general, if $\phi_2(\cdot) = a\phi_1(\cdot) + b$, we say that the function ϕ_2 is *an affine-linear transformation* of ϕ_1 .

Systems of Linear Equations: Infinitely Many Solutions

Some systems of linear equations may have infinitely many solutions. Typically, this happens if we have more variables than equations. For example, the single equation in two variables

$$x - y = 2 \tag{10.15}$$

has infinitely many solutions: for any real number z one has a solution $y = z$ and $x = z + 2$. In this case, we call z a *parameter*.

Some systems have so many solutions that it is convenient to use more than one parameter to write them all down. For example, let us solve the system

$$x_2 - x_1 = x_3 - x_2 = x_4 - x_3 = \cdots = x_t - x_{t-1}, \tag{10.16}$$

which has t variables and $t - 2$ equations. For any numbers n_1 and n_2 , there is a solution with $x_1 = n_1$ and $x_2 - x_1 = n_2$. Then $x_2 = n_1 + n_2$, $x_3 = x_2 + n_2 = n_1 + 2n_2$, and so on,

$$x_i = n_1 + (i - 1)n_2, \dots, i = 1, 2, \dots, t.$$

Such a sequence x_1, x_2, \dots, x_t is called an *arithmetic progression* of length t , with initial term n_1 , and difference n_2 .

The General Formula for Real Solutions

In general, there is a relatively easy theory which, for arbitrary systems of linear equations

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1t}x_t + b_1 = 0 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2t}x_t + b_2 = 0 \\ \dots \\ a_{k1}x_1 + a_{k2}x_2 + \dots + a_{kt}x_t + b_k = 0, \end{cases} \quad (10.17)$$

where a_{ij} and b_i are some coefficients, determines whether it has no solutions, a unique solution, or infinitely many solutions. If there are infinitely many solutions, they can be written down in the form

$$\begin{cases} x_1 = c_{11}n_1 + c_{12}n_2 + \dots + c_{1d}n_d + b'_1 \\ x_2 = c_{21}n_1 + c_{22}n_2 + \dots + c_{2d}n_d + b'_2 \\ \dots \\ x_t = c_{t1}n_1 + c_{t2}n_2 + \dots + c_{td}n_d + b'_d, \end{cases} \quad (10.18)$$

where c_{ij} and b'_i are coefficients, and n_1, n_2, \dots, n_d are parameters. With the notation

$$\psi_i(n_1, n_2, \dots, n_d) = c_{i1}n_1 + c_{i2}n_2 + \dots + c_{id}n_d + b'_i, \quad i = 1, 2, \dots, t, \quad (10.19)$$

(10.18) simplifies to

$$x_i = \psi_i(n_1, n_2, \dots, n_d), \quad i = 1, 2, \dots, t. \quad (10.20)$$

Functions of the form (10.19) are called *affine-linear forms* in d variables. We will denote by $\Psi = (\psi_1, \dots, \psi_t)$ the sequence of t affine-linear forms.

Looking for Integer Solutions

Now, let us assume that all coefficients in (10.17) are integers and we aim to find only *integer* solutions x_1, x_2, \dots, x_t . Such solutions do not always exist: for example, the equation $2x - 2y = 1$ has infinitely many real solutions but no integer solutions because the left-hand side is always even.

Another example: real solutions to the equation

$$x - 2y = 1 \quad (10.21)$$

can be written as a one-parameter family $x = n_1$, $y = \frac{n_1-1}{2}$, but in this case we get non-integer values of y for even values of n_1 . In order for y to be an integer, n_1 must be odd, that is, $n_1 = 2n'_1 + 1$ for some integer n'_1 . Then $x = 2n'_1 + 1$, and $y = \frac{(2n'_1+1)-1}{2} = n'_1$. In general, if system (10.17) with integer coefficients has infinitely many integer solutions, then (possibly after some changes of parameters as above), these solutions can be written in the form (10.20), in such a way that all coefficients in (10.19) are integers.

Looking for *Positive* Integer Solutions

If there are infinitely many integer solutions (x_1, x_2, \dots, x_t) of system (10.17), how many of them are such that all x_i , $i = 1, 2, \dots, t$ are *positive* integers? Various cases are possible. For example, the equation $x + y = 0$ has infinitely many integer solutions, but none of them are such that $x > 0$ and $y > 0$. The equation $x + y = 2$ has infinitely many integer solutions, but only one solution ($x = y = 1$) in positive integers. The system

$$x + y = z - w = 2$$

has infinitely many solutions in positive integers: for any positive integer n it has a solution $x = y = 1$, $z = n + 2$, $w = n$. However, intuitively, this one-parameter family of positive integer solutions is much “smaller” than the two-parameter family

$$x = 2 - n_1, \quad y = n_1, \quad z = n_2 + 2, \quad w = n_2,$$

of all integer solutions to the same system. To say this more formally, let us select a large but finite positive integer N , and assume that $-N \leq n_1 \leq N$ and $-N \leq n_2 \leq N$. Such parameters lead to $(2N + 1)^2$ integer solutions (x, y, z, w) to our system. Among these $(2N + 1)^2$ solutions, only N (those corresponding to $n_1 = 1$, $1 \leq n_2 \leq N$) are such that all x, y, z, w are positive. For large N , $(2N + 1)^2$ is *much* larger than N . Formally,

$$\lim_{N \rightarrow \infty} \frac{N}{(2N + 1)^2} = 0.$$

In contrast, integer solutions of the equation

$$x + y - z = 0$$

are given by the 2-parameter family

$$x = n_1, \quad y = n_2, \quad z = n_1 + n_2,$$

and a “significant portion” of these solutions are such that x, y, z are all positive integers. More formally, in the range $-N \leq n_1 \leq N$ and $-N \leq n_2 \leq N$ there are $(2N+1)^2$ integer solutions, and N^2 of them (corresponding to $1 \leq n_1 \leq N$ and $1 \leq n_2 \leq N$) are positive integers. Now

$$\lim_{N \rightarrow \infty} \frac{N^2}{(2N+1)^2} = \frac{1}{4},$$

hence we can say that 25% of all integer solutions of this equation are positive integers.

In general, for any sequence $\Psi = (\psi_1, \dots, \psi_t)$ of affine-linear forms in d variables with integer coefficients, let $f_\Psi(N)$ be the number of choices of n_1, \dots, n_d in the range $-N \leq n_i \leq N$, $i = 1, 2, \dots, d$ such that all ψ_j , $j = 1, \dots, t$ are positive. Then the constant

$$C_\infty(\Psi) = \lim_{N \rightarrow \infty} \frac{f_\Psi(N)}{(2N+1)^d} \quad (10.22)$$

tells us “how often” all ψ_j are positive integers. Equation (10.22) may also be rewritten as

$$f_\Psi(N) = (C_\infty(\Psi) + o(1))(2N+1)^d,$$

where the notation $o(1)$ denotes a function $\varepsilon_\Psi(N)$ such that $\lim_{N \rightarrow \infty} \varepsilon_\Psi(N) = 0$.

Looking for Prime Solutions

The problem becomes much more interesting and extremely difficult if we ask whether system (10.17) has infinitely many solutions x_1, x_2, \dots, x_d such that all x_i , $i = 1, \dots, d$, are prime numbers (we will call such solutions *prime solutions* for short), that is, positive integers with exactly two divisors. This problem is open even for very simple equations. For example, the conjecture that Eq. (10.15) has infinitely many prime solutions is known as the *twin primes conjecture* and it is one of the most famous unsolved problem in the whole of mathematics. The same conjecture for Eq. (10.21) is known as the Sophie Germain conjecture, and is also considered to be a notoriously difficult open problem. The question of whether system (10.16) has infinitely many prime solutions other than those for which $x_1 = x_2 = \dots = x_t$ is equivalent to the question of whether the primes contain infinitely many arithmetic progressions. This question was resolved positively only in 2008, and this result, discussed in Sect. 8.2, is considered to be one of the greatest achievements in mathematics of the 21st century. Prime solutions to the system

$$x_2 - x_1 = x_3 - x_2 = x_4 - x_3 = \dots = x_t - x_{t-1} = x_{t+1} - 1$$

correspond to arithmetic progressions of length t , consisting of primes whose difference is equal to $p - 1$ for another prime p . Are there infinitely many such progressions?

When are the Values of t Affine-Linear Forms Simultaneously Prime?

Because solutions to (10.17) are given by (10.20), all the questions in the previous section are special cases of the following problem: given a sequence $\Psi = (\psi_1, \dots, \psi_t)$ of affine-linear forms in d variables as in (10.19), are there infinitely many values of n_1, \dots, n_d such that all $\psi_i(n_1, n_2, \dots, n_d)$, $i = 1, 2, \dots, t$ are simultaneously prime? For example, the twin primes conjecture is an instance of this problem with the sequence $(n_1, n_1 + 2)$, the Sophie Germain conjecture corresponds to $(n_1, 2n_1 + 1)$, arithmetic progressions in primes corresponds to the sequence $(n_1, n_1 + n_2, n_1 + 2n_2, \dots, n_1 + (t - 1)n_2)$, while arithmetic progressions in primes whose difference is $p - 1$ for prime p corresponds to the sequence $(n_1, n_1 + n_2, n_1 + 2n_2, \dots, n_1 + (t - 1)n_2, n_2 + 1)$.

Of course, the answer to the question in the previous paragraph is not always positive. For example, there is only one value of n_1 (namely $n_1 = 2$) such that n_1 and $n_1 + 1$ are simultaneously prime. The reason is that, for any integer n_1 , either n_1 or $n_1 + 1$ must necessarily be even, and the only even prime is 2. Similarly, there is only one value of n_1 (namely $n_1 = 3$) such that $n_1, n_1 + 2$, and $n_1 + 4$ are simultaneously prime. This is because one of these numbers is always divisible by 3. Let us call a sequence Ψ *admissible* if no such “obstructions” occur: that is, if for every fixed prime p there exist values of n_1, n_2, \dots, n_d such that all $\psi_i(n_1, n_2, \dots, n_d)$, $i = 1, 2, \dots, t$, are *not* divisible by p .

Another problem occurs with sequences Ψ such as $(n_1, -n_1 - 2)$: for every n_1 , either n_1 or $-n_1 - 2$ is negative, hence they cannot be simultaneously prime. To avoid this problem, we assume that $C_\infty(\Psi) > 0$, where $C_\infty(\Psi)$ is defined in (10.22).

So, if Ψ is admissible and $C_\infty(\Psi) > 0$, can we guarantee that there are infinitely many values of n_1, \dots, n_d such that all $\psi_i(n_1, n_2, \dots, n_d)$, $i = 1, 2, \dots, t$, are simultaneously prime? Moreover, one may also ask *how often* this happens. That is, for positive integer N , let $g_\Psi(N)$ be the number of tuples (n_1, n_2, \dots, n_d) such that $-N \leq n_i \leq N$, $i = 1, 2, \dots, d$ and all ψ_1, \dots, ψ_t are primes. For example, Fig. 10.8 demonstrates that $g_\Psi(12) = 16$ for $\Psi = (n_1, n_1 + n_2, n_1 + 2n_2)$. How fast does $g_\Psi(N)$ grow as a function of N ?

A Significant Advance

The following theorem of Green and Tao [176] solves the described problem for all admissible sequences Ψ with $C_\infty(\Psi) > 0$, which satisfy one additional condition: no ϕ_i is an affine-linear transformation of some ϕ_j .

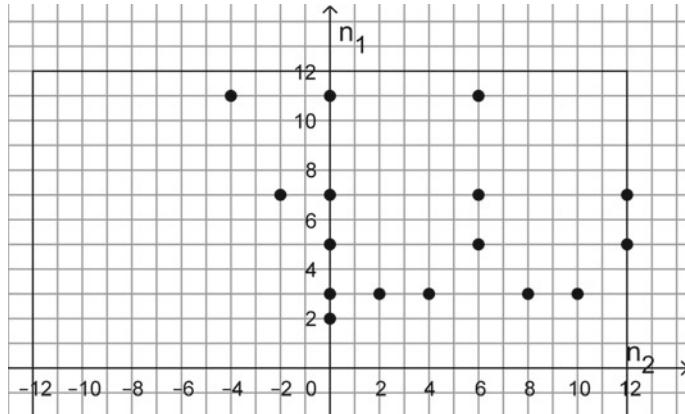


Fig. 10.8 Points (n_1, n_2) in the region $-12 \leq n_i \leq 12$, $i = 1, 2$, such that $n_1, n_1 + n_2, n_1 + 2n_2$ are simultaneously prime

Theorem 10.8 Let $\Psi = (\psi_1, \dots, \psi_t)$ be a sequence of non-constant affine-linear forms in d variables with integer coefficients, see (10.19). Assume that Ψ is admissible, $C_\infty(\Psi) > 0$, and that there are no ψ_i, ψ_j in the sequence such that $\psi_i(\cdot) = a\psi_j(\cdot) + b$ for some constants a, b . Then there are infinitely many values of n_1, \dots, n_d such that all $\psi_i(n_1, n_2, \dots, n_d)$, $i = 1, 2, \dots, t$, are simultaneously prime. Moreover

$$g_\Psi(N) = (C_\Psi + o(1)) \frac{N^d}{\ln^t N},$$

where $C_\Psi > 0$ is a positive constant depending on Ψ .

Theorem 10.8 was proved in [176], assuming the truth of two conjectures. However, Green and Tao then proved these conjectures in later works [177, 178], so that the result is now unconditional.

Applying Theorem 10.8 to the sequence $\Psi = (n_1, n_1 + n_2, n_1 + 2n_2, \dots, n_1 + (t-1)n_2)$, Green and Tao derived that, for large N , there are about $C_t \frac{N^2}{\ln^t N}$ arithmetic progressions $p_1 < p_2 < \dots < p_t < N$ such that all p_i are primes. Here, C_t is a constant which can be explicitly estimated, for example, $C_4 \approx 0.4764$. They also derived similar formulas for the number of such progressions whose difference is $p - 1$ for some prime p , and a lot of other interesting corollaries.

However, Theorem 10.8, despite being a nice and general result, is not applicable to the most difficult problems in the field, e.g. to the twin primes conjecture and the Sophie Germain conjecture. This is because of the additional condition that “no ϕ_i is an affine-linear transformation of some ϕ_j ”. While this condition holds for the sequence $\Psi = (n_1, n_1 + n_2, n_1 + 2n_2, \dots, n_1 + (t-1)n_2)$, corresponding to the arithmetic progressions, and for many other sequences of interest, it fails for sequences $\Psi = (n_1, n_1 + 2)$ and $\Psi = (n_1, 2n_1 + 1)$, which correspond to the twin prime conjecture and the Sophie Germain conjecture, respectively.

Reference

B. Green and T. Tao, Linear equations in primes, *Annals of Mathematics* **171**-3, (2010), 1753–1850.

10.9 Entropies of Multidimensional Shifts of Finite Type may be Impossible to Compute

Real Numbers Which We Can Compute

Can you compute the length of the diagonal of the unit square, or the circumference of the circle of unit diameter? If the length of the diagonal is x , Pythagoras' theorem implies that $x^2 = 1^2 + 1^2 = 2$, hence $x = \sqrt{2}$. It is known that $\sqrt{2}$ is an irrational number, that is, cannot be represented as $\frac{a}{b}$ for some integers a and b . The decimal expansion of $\sqrt{2}$ starts with

$$\sqrt{2} = 1.41421356237309504880\dots$$

and continues up to infinity. There is no point in talking about “computing” such a number exactly. The best we can do is to *approximate* it within any given precision. For $\sqrt{2}$, a simple and fast approximation procedure is the following one. Start with $a_0 = 1$, and compute

$$a_{n+1} = \frac{a_n}{2} + \frac{1}{a_n}.$$

Then $a_1 = \frac{1}{2} + \frac{1}{1} = \frac{3}{2} = 1.5$, $a_2 = \frac{(3/2)}{2} + \frac{1}{(3/2)} = \frac{17}{12} \approx 1.416$. Continuing in this way, we get $a_3 = \frac{577}{408} \approx 1.414215$, $a_4 = \frac{665857}{470832} \approx 1.4142135623746$, and so on. As you can see, a_1, a_2, a_3 , and a_4 approximate $\sqrt{2}$ to within 1, 3, 6, and 12 decimal digits, respectively. One may prove that, computing this sequence further and further, we can approximate $\sqrt{2}$ better and better, with the number of correct digits approximately doubling after each iteration. After just 30 iterations, we can compute hundreds of millions of digits of $\sqrt{2}$ correctly.

The circumference of the circle of unit diameter is known as the number π , whose decimal expansion starts with

$$\pi = 3.14159265358979323846\dots$$

This is also an irrational number, and computing its digits was an old problem in mathematics. In 1910, Ramanujan discovered a formula which gives (approximately) eight new correct digits of π after each iteration. With 21st century computers, millions of millions of digits of π have currently been computed.

In general, a real number x is called *computable* if we can approximate it to within any given accuracy. More formally, x is computable if there exists a computer

program which, for a given integer n , returns integers a_n and b_n such that $\left| x - \frac{a_n}{b_n} \right| < \frac{1}{n}$.

Real Numbers Which We Cannot Compute

The discussion above implies that the real numbers $\sqrt{2}$ and π are computable. One may ask: are all real numbers computable? Intuition from physics and other disciplines seems to suggest that the answer should be “Yes”. After all, physical constants such as the speed of light are real numbers, and it is hard to imagine some physical constant which cannot be estimated even approximately. However, relatively easy mathematics shows that the answer is in fact “No”—there are real numbers which are not computable!

Why so? Well, by contradiction, imagine that for any real number x there is a computer program, written in some programming language, which computes x . In fact, any computer program is just a finite set of symbols in some alphabet. We can sort all computer programs by length (shorter goes first), arrange programs of equal lengths in alphabetical order, and then enumerate all these programs: the first one, the second one, and so on. Let x_n be the real number computed by program number n . Let $x^* \in (0, 1)$ be the real number whose n -th decimal digit after the dot is the same as the n -th decimal digit of $x_n + 10^{-n}$. Then $x^* \neq x_n$ for all n . Hence, x^* is not computable by any computer program.

We have just proved that there are real numbers x which are not computable: there is no program or procedure which computes digits of x as far as we want, no program returning a sequence of rational numbers $\frac{a_n}{b_n}$ such that $\frac{a_n}{b_n} - \frac{1}{n} < x < \frac{a_n}{b_n} + \frac{1}{n}$ for all n . The last condition says that we cannot approximate x at the same time from above and from below. In some applications, it suffices to approximate x from above only. Formally, a real number x is called *right recursively enumerable* if there exists a computer program which, for a given integer n , returns integers a_n and b_n such that $x \leq \frac{a_n}{b_n}$ and $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = x$. In other words, we can have at least upper bounds for x of higher and higher accuracy. It is known that there are real numbers x which are not computable but right recursively enumerable. However, the same proof as above implies that there are also real numbers which are not even right recursively enumerable.

A Shift of Finite Type (SFT) Model: First Example

The proof above implies that there are some real numbers x which we cannot compute, even approximately. But could it be that these are just some artificially constructed “uninteresting” numbers which do not arise in any application? Unfortunately, as we will see below, this is not the case: some non-computable numbers really arise in applications in a natural way.

When studying a wide variety of physical systems, one often discovers that the behaviour of the systems are determined by laws which are local in nature, and may “forbid” the occurrence of some “local configurations”. For example, repulsing electrical forces forbid particles with the same sign of electric charge from being too “close” to each other. The question is then to study the global properties of systems with such “forbidden local configurations”.

An important mathematical model used to describe such phenomena is called “A shift of finite type (SFT) model”. To start, imagine that we need to colour points on the coordinate line with integer coordinates $\dots - 2, -1, 0, 1, 2, \dots$ in two colours, black and white, in such a way that no two adjacent points n and $n + 1$ are coloured in the same way. In other words, points $(n, n + 1)$ can be either (black, white) or (white, black), but not (black, black) or (white, white). In this model, black and white coloured points may model negatively and positively charged particles, respectively, and the forbidden (black, black) and (white, white) configurations arise as a consequence of repulsing electrical forces. However, this particular model is too restrictive, and allows only two configurations: either points with even coordinates are white and points with odd coordinates are black, or vice versa.

An SFT Model: Further Examples

A more interesting example is if the allowed patterns for points $(n, n + 1)$ are (black, white), (white, black), and (white, white)—that is, only the (black, black) configuration is forbidden. Such rules may arise if, for example, black and white coloured points model negatively charged and *neutral* particles, respectively. In this case, infinitely many different colourings are possible. For example, any three consecutive points can then be coloured as BWB (here and below B denotes black and W denotes white), BWW, WBW, WWB, or WWW—5 ways in total. Further, four consecutive points can then be coloured as BWBW, BWWB, WBWB, BWWW, WBWW, WWBW, WWWB, or WWWW—8 ways. In general, let X be the set of all colourings avoiding the (black, black) pattern, and let $f_X(n)$ be the number of possible colourings of n consecutive points. We have just seen that $f_X(3) = 5$ and $f_X(4) = 8$. For large n , it is known that $f_X(n)$ is approximately equal to $2^{h(X)n}$ for some constant $h(X)$. The quantity $h(X)$ controls how “large” the set X of possible colourings is, and is known as the (topological) entropy of X . From equation $f_X(n) \approx 2^{h(X)n}$ we can informally deduce that $h(X) \approx \frac{1}{n} \log_2 f_X(n)$. Formally, $h(X)$ is defined as the limit $h(X) = \lim_{n \rightarrow \infty} \frac{1}{n} \log_2 f_X(n)$. This quantity characterizes the “flexibility” or the “level of disorder” in the physical system we are trying to model. In our particular case, one can prove that $h(X) = \log_2 \left(\frac{1+\sqrt{5}}{2} \right) \approx 0.694$.

As another example, let us consider colourings of all points with integer coordinates (k, m) in the *plane*, and let X be a set of colourings such that no three points with coordinates (k, m) , $(k, m + 1)$, and $(k + 1, m)$ (they form a kind of “corner”) all have the same colour. That is, out of 8 colourings of the “corner” 2 (all points white and all points black) are forbidden and the other 6 are allowed, see Fig. 10.9a.

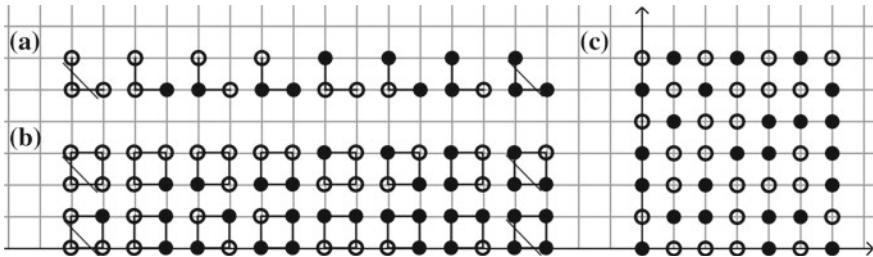


Fig. 10.9 Possible and forbidden colourings of the “corder”, 2×2 square, and 7×7 square

A physical interpretation of this model may be that we are studying particles moving along a 2-dimensional surface, black and white colours correspond to negative and positive charges, but this time repulsing electrical forces are weaker, and forbid only configurations with *three* particles of the same colour to be close to each other.

In such a system, there are 12 possible (and 4 “impossible”) colourings of any “ 2×2 square”, that is, points with coordinates (k, m) , $(k, m + 1)$, $(k + 1, m)$, and $(k + 1, m + 1)$, see Fig. 10.9b. More generally, we can count the number $f_X(n)$ of possible colourings of an “ $n \times n$ square” consisting of n^2 points of the form $(k + i, m + j)$, $0 \leq i \leq n - 1$, $0 \leq j \leq n - 1$ (an example of such a colouring for $n = 7$ is presented in Fig. 10.9c), and calculate the entropy using formula $h(X) = \lim_{n \rightarrow \infty} \frac{1}{n^2} \log_2 f_X(n)$.

An SFT Model: General Definition

In reality, there are more than two types of particles, so we may need to use more than two colours in our model. Also, particles are usually located in 3-dimensional space. Moreover, if we take into account velocities, then each particle may be described by 6 coordinates: 3 describes space location and 3 others, the velocity vector. So, in general, we need to consider colourings of integer points \mathbb{Z}^d of d -dimensional space using any finite number of colours.

In the example above, a corner “ (k, m) , $(k, m + 1)$, $(k + 1, m)$ ” is just a set of three fixed points “ $(0, 0)$, $(0, 1)$, $(1, 0)$ ”, translated by a vector (k, m) . In general, let $F \subset \mathbb{Z}^d$ be any finite set of points, and, for any $x \in \mathbb{Z}^d$, by the “translate of F by x ” we mean the set $\{x + y, | y \in F\}$. Further, let L be a collection of colourings of F . We interpret L as a set of “allowed” colourings of F and its translates, and all the other colourings of F and its translates are “forbidden”. A colouring of \mathbb{Z}^d is called *admissible* for L if all translates of F are coloured according to a pattern belonging to L . In other words, no translate of F is coloured in a forbidden way. The *shift of finite type (SFT)* defined by L is the set X of all admissible colourings of \mathbb{Z}^d .

For every positive integer n , let F_n be the set of points $(y_1, \dots, y_d) \in \mathbb{Z}^d$ such that $1 \leq y_i \leq n$ for all $i = 1, 2, \dots, d$. For the SFT X , let $f_X(n)$ be the number of different colourings of F_n which can appear in some colouring $x \in X$. The (topological) entropy of X is then defined as

$$h(X) = \lim_{n \rightarrow \infty} \frac{1}{n^d} \log_2 f_X(n).$$

Can We Compute the Entropy in an SFT Model?

In applications, it is important to *compute* the real number $h(X)$ for a given SFT X . Of course, we can always compute $\frac{1}{n^d} \log_2 f_X(n)$ for any fixed n , and it is known that $h(X) \leq \frac{1}{n^d} \log_2 f_X(n)$ for all n . Hence, for any SFT X , the real number $h(X)$ is right recursively enumerable. But is it computable? The following theorem, proved in [203], implies a negative answer.

Theorem 10.9 *Let $d \geq 2$. A real number y is right recursively enumerable if and only if $y = h(X)$ for some d -dimensional SFT X .*

As we have just explained, the “if” direction in Theorem 10.9 is easy. The really interesting and difficult direction is the “only if” part, stating that, for *every* right recursively enumerable y , there is an SFT X such that $y = h(X)$. Because some right recursively enumerable real numbers are not computable, Theorem 10.9 implies that, for some SFT X , there is no algorithm to compute its entropy $h(X)$. Not only is there no algorithm which takes the description of an SFT as an input and produces its entropy as an output, there isn’t even an algorithm which computes $h(X)$ as an individual real number.

Reference

M. Hochman and T. Meyerovitch, A characterization of the entropies of multidimensional shifts of finite type, *Annals of Mathematics* **171**-3, (2010), 2011–2038.

10.10 Subgroups of 2-Generated Groups

Addition, Multiplication, and the Notion of a Group

The addition operation for integers satisfy the following properties: (i) the sum of any two integers is again an integer; (ii) $(a + b) + c = a + (b + c)$ for all integers a, b, c ; (iii) there exists a special integer 0 such that $a + 0 = 0 + a = a$ for all integers a ; and (iv) for every integer a , there exists an integer b (namely, $b = -a$) such that $a + b = b + a = 0$.

In fact, the addition operation for real numbers, or, for example, for functions $f : \mathbb{R} \rightarrow \mathbb{R}$, where \mathbb{R} is the set of real numbers, satisfies the same properties. Moreover, similar properties hold for the set of all non-zero real numbers with the *multiplication* operation \cdot , namely, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$, there is a special real number 1 such that $a \cdot 1 = 1 \cdot a = a$ for all a , and for every $a \neq 0$, there exists a b , (namely, $b = 1/a$), such that $a \cdot b = b \cdot a = 1$.

In fact, examples of operations satisfying these properties can be found in essentially all areas of mathematics, see Sect. 1.5 for more examples. To study all such examples at once, mathematicians introduced the notion of a *group*. A group is any set G and operation \cdot which assigns to any two elements $a, b \in G$ a third one, denoted $a \cdot b$, or just ab , such that the following properties hold.

- (i) $a \cdot b \in G$ for all $a, b \in G$;
- (ii) $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all $a, b, c \in G$;
- (iii) there exists an $e \in G$ (with is called the *identity element*) such that $a \cdot e = e \cdot a = a$ for all $a \in G$;
- (iv) for every $a \in G$, there exists an element $b \in G$ (also denoted a^{-1}) such that $a \cdot b = b \cdot a = e$.

For example, the set \mathbb{Z} of all integers is a group with operation $+$, while the set $\mathbb{R} \setminus \{0\}$ of all non-zero real numbers is a group with operation \cdot .

Some Examples of Finite Groups

A group G is called *finite* if G is a finite set. The simplest group is just a one-element set $G = \{e\}$ with operation $e \cdot e = e$. This group is called the *trivial group*. The next simplest group is the 2-element group $G = \{e, f\}$ with operations $e \cdot e = f \cdot f = e$ and $e \cdot f = f \cdot e = f$. A standard notation for this group is $\mathbb{Z}/2\mathbb{Z}$.

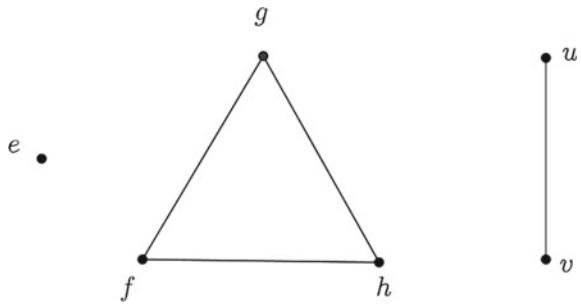
Another example of a finite group is a permutation group. For example, given 3 objects a, b, c , there are 6 ways to “permute” them: $abc \rightarrow abc$ (no change), $abc \rightarrow bac$ (exchange the order of the first two objects), $abc \rightarrow cba$ (exchange the first one with the third one), $abc \rightarrow acb$ (exchange the order of the last two objects), $abc \rightarrow bca$ (cyclic shift left), and $abc \rightarrow cab$ (cyclic shift right). For any two permutations f and g , their *composition*, denoted as $f \circ g$, is a permutation resulting from applying f after g . For example, let us compute $f \circ g$ for $f = (abc \rightarrow acb)$ and $g = (abc \rightarrow bac)$. For this, we start with abc , apply g (that is, exchange the order of the first two objects) to get bac , and then apply f (that is, exchange the order of the last two objects) to get bca . Hence, $(abc \rightarrow acb) \circ (abc \rightarrow bac) = (abc \rightarrow bca)$. One may check that the operation \circ satisfies the properties (i)–(iv) above, hence the set of all permutations is a group with this operation. This group is usually denoted by S_3 .

Conjugacy Classes

For any group G , elements $f \in G$ and $h \in G$ are called *conjugate* to each other if there exists an element $g \in G$ such that $f \cdot g = g \cdot h$. For example, permutations $f = (abc \rightarrow acb)$ and $h = (abc \rightarrow cba)$ are conjugate in S_3 because, for $g = (abc \rightarrow bac)$, we have

$$f \cdot g = (abc \rightarrow acb) \circ (abc \rightarrow bac) = (abc \rightarrow bca)$$

Fig. 10.10 Three conjugacy classes of S_3 : elements of S_3 are points, points corresponding to conjugate elements are connected



and

$$g \cdot h = (abc \rightarrow bac) \cdot (abc \rightarrow cba) = (abc \rightarrow bca).$$

In fact, one can check that all three permutations f, g, h in this example are conjugate to each other. The cyclic shifts $u = (abc \rightarrow bca)$ and $v = (abc \rightarrow cab)$ are also conjugate to each other, while the “no shift” permutation $e = (abc \rightarrow abc)$ is not conjugate to any other element. Hence, the whole group S_3 can be partitioned into three disjoint subsets $S_3 = \{e\} \cap \{f, g, h\} \cap \{u, v\}$ such that (i) every two elements from the same subset are conjugate to each other, and (ii) every two elements from different subsets are *not* conjugate, see Fig. 10.10. In fact, *any* group G can be partitioned into subsets with these properties, and these subsets are called *conjugacy classes* of G .

The trivial group $G = \{e\}$ has just one conjugacy class. In fact, this is the only group with one conjugacy class. The reason is that, in any group G , the identity element e is not conjugate to any element $f \neq e$. Indeed, if it were, we would have $e \cdot g = g \cdot f$ for some $g \in G$, which implies that $g = g \cdot f$. Multiplying both sides by g^{-1} , we get $g^{-1} \cdot g = g^{-1} \cdot g \cdot f$, or $e = f$, a contradiction.

Elements of Finite and Infinite Order

Let us select any non-identity element a in any group G . Let us denote by $a^2 = a \cdot a$ its product with itself, $a^3 = a^2 \cdot a = a \cdot a \cdot a$, and so on, and write down the infinite sequence

$$a, a^2, a^3, \dots$$

There are two cases possible. Case 1 is that this sequence contains repetitions, that is, $a^i = a^j$ for some $i < j$. Multiplying both sides of this equality by a^{-1} i times, we get $e = a^{j-i}$. The lowest positive integer n such that $a^n = e$ is called the *order* of a in G . For example, $f \circ f = e$ for permutation $f = (abc \rightarrow acb)$ in S_3 , hence f has order 2. Also, $u \circ u \neq e$ but $u \circ u \circ u = e$ for $u = (abc \rightarrow bca)$, hence this permutation has order 3.

Case 2, possible only if G is an infinite group, is that all a^n in this sequence are different. For example, this happens if $G = \mathbb{R} \setminus \{0\}$ with multiplication operation, and $a = 2$. Then all numbers in the sequence $2, 4, 8, 16, 32, \dots$ are of course different. In this case, we say that the element a has an *infinite* order in G .

A subset H of a group G is called a *subgroup* of G if (i) $a \cdot b \in H$ for all $a, b \in H$; (ii) $e \in H$, where e is the identity element of G ; and $a^{-1} \in H$ for every $a \in H$. In other words, H is also a group with the same operation \cdot . For example, the set \mathbb{Z} of integers with operation $+$ is a subgroup of the set \mathbb{R} of real numbers with the same operation.

As another example, in any group G , take any $a \in G$, form the sequence a, a^2, a^3, \dots as above, and also denote by a^0 the identity element e and a^{-n} the inverse element to a^n for $n = 1, 2, \dots$. Then $H = \{a^n, n \in \mathbb{Z}\}$ is a subgroup of G .

If a has infinite order, than all $a^n, n \in \mathbb{Z}$, are different. In this case, H is an infinite sequence $\dots, a^{-2}, a^{-1}, a^0, a^1, a^2, \dots$, which resembles the sequence of integers $\dots, -2, -1, 0, 1, 2, \dots$. In fact, $a^n \cdot a^m = a^{n+m}$, hence to calculate the product of any elements of H , all we need to do is to just add the corresponding integers. In other words, the group H with operation \cdot is just the group \mathbb{Z} of integers with operation $+$, written in a different “format”. In such cases we say that the group H is *isomorphic* to the group \mathbb{Z} . Formally, we say that two groups G and H are *isomorphic* if there exists a one-to-one correspondence between their elements which preserves the group operations. In our example, this correspondence is given by $n \leftrightarrow a^n$.

If a group G has a subgroup H isomorphic to a group K , we say that the group K can be *embedded* into the group G . For example, we have just demonstrated that the group \mathbb{Z} of integers with operation $+$ can be embedded into any group G which has at least one element a of infinite order.

Countable, Uncountable, and 2-generated Groups

An infinite set S is called *countable* if we can list all its elements in the form of a sequence

$$s_1, s_2, s_3, \dots$$

Similarly, a group G is countable if we can list all its elements in such a sequence, and it is called *uncountable* otherwise. For example, the group \mathbb{Z} of integers is countable, because we can form a sequence

$$0, 1, -1, 2, -2, 3, -3, \dots$$

which contains all integers. On the other hand, it is known that there is no such sequence containing all real numbers, hence the group \mathbb{R} of real numbers is uncountable.

A group G is called *2-generated* if there exist two elements $a \in G$ and $b \in G$, such that any element $g \in G$ can be written as a product $g = g_1 g_2 \dots g_n$, where each g_i is either a , or b , or a^{-1} , or b^{-1} . In this case, we say that “ a and b generates

G'' . For example, it is easy to prove that every integer g can be written as $g = g_1 + g_2 + \dots + g_n$ where each g_i is either 2 or 3, or -2 , or -3 , hence the group \mathbb{Z} of integers can be generated by $a = 2$ and $b = 3$ (in fact, it can also be generated by one integer $a = 1$). In contrast, the real number $g = 0.5$ cannot be written in such a form, hence the real numbers $a = 2$ and $b = 3$ do not generate the group \mathbb{R} . Moreover, it is easy to check that no other choice of a and b works, so the group \mathbb{R} is not 2-generated. In fact, all uncountable groups are not 2-generated.

The Theorem and Corollaries

For any group G , let $\pi(G)$ be the set of all (finite) positive integers n such that there exists an element $a \in G$ of order n . In 2010, Denis Osin [295] proved the following theorem.

Theorem 10.10 *Any countable group G can be embedded into a 2-generated group C such that any two elements of the same order are conjugate in C and $\pi(G) = \pi(C)$.*

This theorem turned out to be extremely useful, has a number of interesting corollaries, and immediately resolves some important open questions in the field. For example, as we mentioned above, the group $\mathbb{Z}/2\mathbb{Z}$ has exactly 2 conjugacy classes. This group, which has just two elements, is also trivially 2-generated. Until 2010, it was open whether there exists any other 2-generated³ group with exactly two conjugacy classes.

Applying Theorem 10.10 to any group G in which all elements, except the identity element, have infinite order (such groups are called torsion-free), we obtain that *any countable torsion-free group can be embedded into a torsion-free 2-generated group with exactly 2 conjugacy classes*. In turn, this implies that *there exists an uncountable set of pairwise non-isomorphic torsion-free 2-generated groups with exactly 2 conjugacy classes*. Of course, this resolves the open question above positively in a very strong sense. Some other interesting corollaries from Theorem 10.10 are presented in [295] as well.

Reference

D. Osin, Small cancellations over relatively hyperbolic groups and embedding theorems, *Annals of Mathematics* **172**-1, (2010), 1–39.

³More generally, even the existence of a group, except for $\mathbb{Z}/2\mathbb{Z}$, with any finite number of generators and two conjugacy classes was an open question!

10.11 On the Number of Quintic Fields with Bounded Discriminant

Natural Numbers, Integers, Rational Numbers, and Beyond

In ancient times, the first numbers people discovered were natural numbers $0, 1, 2, 3, 4, \dots$, which they used for counting. Later, to be able to do subtractions like $4 - 6$, negative numbers were introduced, which, together with natural numbers, form a set of *integers*. Also, to be able to perform division like $\frac{4}{6}$, people introduced *rational* numbers, that is, numbers of the form $x = \frac{b}{a}$ for integers $a \neq 0$ and b . The set of all rational numbers is usually denoted by \mathbb{Q} . If a and b have a common factor $d > 1$, that is, $a = da'$ and $b = db'$ for integers a' and b' , then the factor d can be cancelled out: $x = \frac{b}{a} = \frac{db'}{da'} = \frac{b'}{a'}$, for example, $\frac{4}{6} = \frac{2 \cdot 2}{2 \cdot 3} = \frac{2}{3}$. If a' and b' has a common factor $d' > 1$, it can be cancelled out as well, and, at the end, any rational number x can be written as an irreducible fraction $\frac{n}{m}$, that is, such that n and m have no common factors greater than 1. The sum, difference, product, and ratio (except for division by 0) of any two rational numbers is again a rational number.

However, people needed some “new” numbers to be able to solve equations. For example, the positive root to the equation $x^2 = 3$, denoted $x = \sqrt{3}$, is not a rational number. If it were, it could be written as an irreducible fraction $x = \frac{n}{m}$. Then $(\frac{n}{m})^2 = 3$ implies that $n^2 = 3m^2$, which means that n is a multiple of 3. Then $n = 3k$ for some integer k , and $(3k)^2 = 3m^2$, or $3k^2 = m^2$, implies that m is also a multiple of 3, contradicting the assumption that the fraction $\frac{n}{m}$ is irreducible. Hence, $\sqrt{3}$ is not a rational number.

“Adding” $\sqrt{3}$ to the Set of Rational Numbers

If we have introduced negative numbers to be able to perform subtraction, and have introduced fractions to be able to perform division, can we “add” $\sqrt{3}$ to the set \mathbb{Q} of rational numbers to be able to solve the equation $x^2 = 3$? In this case, we also need to “add” all numbers of the form $b\sqrt{3}$, $b \in \mathbb{Q}$, to be able to perform multiplication by our “new” number $\sqrt{3}$, and, more generally, all numbers of the form $a + b\sqrt{3}$ for rational a, b , to be able to perform addition. Let $\mathbb{Q}(\sqrt{3}) = \{a + b\sqrt{3}, |a, b \in \mathbb{Q}\}$ be the resulting set we get. If $x = a + b\sqrt{3} \in \mathbb{Q}(\sqrt{3})$ and $y = c + d\sqrt{3} \in \mathbb{Q}(\sqrt{3})$, then so is their sum $x + y = (a + c) + (b + d)\sqrt{3}$, difference $x - y = (a - c) + (b - d)\sqrt{3}$, and product $x \cdot y = (a + b\sqrt{3}) \cdot (c + d\sqrt{3}) = (ac + 3bd) + (ad + bc)\sqrt{3}$. Also, for every non-zero $y = a + b\sqrt{3} \in \mathbb{Q}(\sqrt{3})$,

$$\frac{1}{y} = \frac{1}{a + b\sqrt{3}} = \frac{a - b\sqrt{3}}{(a + b\sqrt{3})(a - b\sqrt{3})} = \frac{a}{a^2 - 3b^2} + \frac{-b}{a^2 - 3b^2}\sqrt{3} \in \mathbb{Q}(\sqrt{3}).$$

Hence, if x and $y \neq 0$ belong to $\mathbb{Q}(\sqrt{3})$, then so does their ratio $\frac{x}{y} = x \cdot \left(\frac{1}{y}\right)$. Hence, we do not need to add any more numbers to $\mathbb{Q}(\sqrt{3})$ to be able to perform four basic operations.

In the example above, we did not need to “add” any further numbers to be able to multiply a “new” number $\sqrt{3}$ by itself, because $\sqrt{3} \cdot \sqrt{3} = 3$ is rational. However, if we would like to be able to solve the equation $x^3 = 3$ (instead of $x^2 = 3$), and “add” its solution $x = \sqrt[3]{3}$ to the set \mathbb{Q} of rational numbers, we would also need to “add” $x \cdot x = x^2 = \sqrt[3]{9}$ to be able to perform multiplication, and then we would also need to add all numbers in the form $a + b\sqrt[3]{3} + c\sqrt[3]{9}$ for rational a, b, c . This set is denoted by $\mathbb{Q}(\sqrt[3]{3})$.

Algebraic Numbers and Number Fields

In general, any polynomial equation with rational coefficients can, after division by the leading coefficient, be written in the form

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 = 0, \quad (10.23)$$

where a_0, a_1, \dots, a_{n-1} are rational. The solutions of such equations are called *algebraic numbers*. For example, $\sqrt{3}$ and $\sqrt[3]{3}$ are solutions to the equations $x^2 - 3 = 0$ and $x^3 - 3 = 0$, respectively, hence they are algebraic numbers.

In fact, any algebraic number γ is a solution to *many* equations of the form (10.23). For example, $\sqrt[3]{3}$ solves not only the equation $x^3 - 3 = 0$, but also, for example, the equation $x^8 - 3x^5 = 0$. The minimal n in (10.23) such that γ is a solution to (10.23) is called the *degree* of an algebraic number γ . For example, the degree of $\sqrt[3]{3}$ is 3, while the degree of $\sqrt{3}$ is 2.

For any algebraic number γ of degree n , the set of all numbers x of the form

$$x = b_0 + b_1\gamma + b_2\gamma^2 + \dots + b_{n-1}\gamma^{n-1},$$

where b_0, b_1, \dots, b_{n-1} are some rational numbers, is denoted by $\mathbb{Q}(\gamma)$ and called a *number field* of degree n . One may show that if x and y belongs to $\mathbb{Q}(\gamma)$, then so do $x + y$, $x - y$, $x \cdot y$, and also the ratio x/y , provided that $y \neq 0$. The sets $\mathbb{Q}(\sqrt{3}) = \{a + b\sqrt{3}, |a, b \in \mathbb{Q}\}$ and $\mathbb{Q}(\sqrt[3]{3}) = \{a + b\sqrt[3]{3} + c\sqrt[3]{9}, |a, b, c \in \mathbb{Q}\}$ described above are examples of number fields.

Bases of a Number Field

Now, let γ^* be a positive solution to the equation $x^2 - 2x - 11 = 0$. Then γ^* is an algebraic number of degree 2, and $\mathbb{Q}(\gamma^*) = \{c + d\gamma^*, |c, d \in \mathbb{Q}\}$ is, by definition, a number field. However, in fact $\gamma^* = \frac{-(-2) + \sqrt{(-2)^2 - 4 \cdot 1 \cdot (-11)}}{2} = 1 + 2\sqrt{3}$. Hence,

$$c + d\gamma^* = c + d(1 + 2\sqrt{3}) = (c + d) + (2d)\sqrt{3},$$

which is the same as $a + b\sqrt{3}$ with $a = c + d$ and $b = 2d$. In other words, the field $\mathbb{Q}(\gamma^*)$ is the “same” field as $\mathbb{Q}(\sqrt{3})$, just written in a different notation. Such fields are called *isomorphic*.

Hence, each element $x = a + b\sqrt{3}$ of the number field $\mathbb{Q}(\sqrt{3})$ can also be represented as $x = c + d\gamma^*$ for some $c, d \in \mathbb{Q}$. In general, we say that n elements e_1, e_2, \dots, e_n of a number field $\mathbb{Q}(\gamma)$ form a *basis* of $\mathbb{Q}(\gamma)$ if every $x \in \mathbb{Q}(\gamma)$ can be written as

$$x = c_1e_1 + c_2e_2 + \dots + c_ne_n \quad (10.24)$$

for some rational numbers c_1, c_2, \dots, c_n . The definition of a number field implies that $1, \gamma, \gamma^2, \dots, \gamma^{n-1}$ always forms a basis of $\mathbb{Q}(\gamma)$. However, this basis is not unique. For example, as we saw above, the number field $\mathbb{Q}(\sqrt{3})$ has basis $1, \sqrt{3}$, but also basis $1, 1 + 2\sqrt{3}$.

The Trace of an Element in a Number Field

Let us fix any number field $\mathbb{Q}(\gamma)$, any $x \in \mathbb{Q}(\gamma)$, and any basis e_1, e_2, \dots, e_n of $\mathbb{Q}(\gamma)$. Then the n numbers $x \cdot e_1, x \cdot e_2, \dots, x \cdot e_n$ belong to $\mathbb{Q}(\gamma)$, and can therefore be written in the form (10.24) with some coefficients:

$$x \cdot e_i = c_{i1}e_1 + c_{i2}e_2 + \dots + c_{in}e_n, \quad i = 1, 2, \dots, n.$$

Then the *trace* of x in $\mathbb{Q}(\gamma)$, denoted $\text{Tr}(x)$, is the sum $c_{11} + c_{22} + \dots + c_{nn}$. For example, consider $x = 1 + \sqrt{3}$ in $\mathbb{Q}(\sqrt{3})$ and the basis $1, \sqrt{3}$. Then

$$x \cdot e_1 = (1 + \sqrt{3}) \cdot 1 = 1 \cdot 1 + 1 \cdot \sqrt{3},$$

and

$$x \cdot e_2 = (1 + \sqrt{3}) \cdot \sqrt{3} = 3 \cdot 1 + 1 \cdot \sqrt{3},$$

hence $c_{11} = c_{12} = 1$, $c_{21} = 3$, $c_{22} = 1$, and the trace $\text{Tr}(\sqrt{3}) = c_{11} + c_{22} = 2$. Now, consider the same $x = 1 + \sqrt{3}$ but with basis $1, 1 + 2\sqrt{3}$. Then

$$x \cdot e_1 = (1 + \sqrt{3}) \cdot 1 = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot (1 + 2\sqrt{3}),$$

and

$$x \cdot e_2 = (1 + \sqrt{3}) \cdot (1 + 2\sqrt{3}) = 7 + 3\sqrt{3} = \frac{11}{2} \cdot 1 + \frac{3}{2} \cdot (1 + 2\sqrt{3}),$$

hence $c_{11} = c_{12} = 1/2$, $c_{21} = 11/2$, $c_{22} = 3/2$, and the trace $\text{Tr}(\sqrt{3}) = c_{11} + c_{22} = 1/2 + 3/2 = 2$ —the same as with basis $1, \sqrt{3}$! This is not a coincidence and it is true in general—the trace $\text{Tr}(x)$ is the same for all choices of the basis and depends on x only. Using this result, we can always choose the simplest possible basis to calculate the trace. For example, in the field $\mathbb{Q}(\sqrt{3})$, we can use the basis $1, \sqrt{3}$ to conclude that the trace of any $x = a + b\sqrt{3}$ is equal to $2a$.

Algebraic Integers and Integral Bases

A number γ is called an *algebraic integer* if it is a solution of Eq. (10.23) with *integer* coefficients. For example, $\sqrt{3}$ and $1 + 2\sqrt{3}$ are solutions to the equations $x^2 - 3 = 0$ and $x^2 - 2x - 11 = 0$, respectively, hence they are algebraic integers. More generally, for any integers m and n , the number $m + n\sqrt{3}$ is a solution to the equation $x^2 - 2mx + m^2 - 3n^2 = 0$ (check!) with integer coefficients, hence it is an algebraic integer. Conversely, it is not difficult to prove that *all* algebraic integers in $\mathbb{Q}(\sqrt{3})$ can be written as $m + n\sqrt{3}$ for integers m, n .

In general, we say that n elements e_1, e_2, \dots, e_n of a number field $\mathbb{Q}(\gamma)$ form an *integral basis* of $\mathbb{Q}(\gamma)$ if every algebraic integer $x \in \mathbb{Q}(\gamma)$ can be written in the form (10.24) with *integer* coefficients c_1, c_2, \dots, c_n . For example, $1, \sqrt{3}$ is an integral basis of $\mathbb{Q}(\sqrt{3})$, see Fig. 10.11a. On the other hand, the basis $1, 1 + 2\sqrt{3}$ is not an integral basis, because, for example, the algebraic integer $\sqrt{3}$ cannot be written as $\sqrt{3} = m + n(1 + 2\sqrt{3})$ for some integers m and n , see Fig. 10.11b. On the other hand, the pair $1, 1 + \sqrt{3}$ forms an integral basis, because every algebraic integer $m + n\sqrt{3}$ in $\mathbb{Q}(\sqrt{3})$ can be written as $(m - n) \cdot 1 + n \cdot (1 + \sqrt{3})$.

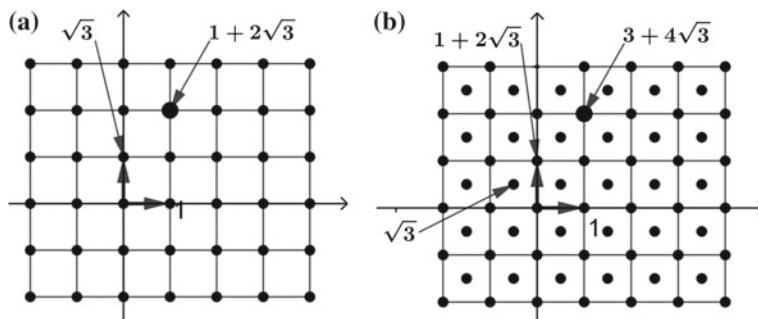


Fig. 10.11 a and b depicts algebraic integers (marked as dots) of $\mathbb{Q}(\sqrt{3})$ using bases $1, \sqrt{3}$ and $1, 1 + 2\sqrt{3}$, respectively

The Integral Trace Form, and the Discriminant of a Number Field

Given an integral basis e_1, e_2, \dots, e_n of a number field $\mathbb{Q}(\gamma)$, we can calculate n^2 numbers $t_{ij} = \text{Tr}(e_i e_j)$, $1 \leq i, j \leq n$, and arrange them into an $n \times n$ table, or *matrix*. This matrix is called the *integral trace form*. For example, for the integral basis $1, \sqrt{3}$ of $\mathbb{Q}(\sqrt{3})$, the integral trace form is $\begin{bmatrix} \text{Tr}(1 \cdot 1) & \text{Tr}(1 \cdot \sqrt{3}) \\ \text{Tr}(\sqrt{3} \cdot 1) & \text{Tr}(1 \cdot \sqrt{3}) \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 6 \end{bmatrix}$. For any 2×2 matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, the number $ad - bc$ is called the *determinant* of this matrix. For example, the determinant of our integral trace form is $2 \cdot 6 - 0 \cdot 0 = 12$. For the integral basis $1, 1 + \sqrt{3}$ of $\mathbb{Q}(\sqrt{3})$, the integral trace form is $\begin{bmatrix} \text{Tr}(1^2) & \text{Tr}(1 \cdot (1 + \sqrt{3})) \\ \text{Tr}((1 + \sqrt{3}) \cdot 1) & \text{Tr}((1 + \sqrt{3})^2) \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 8 \end{bmatrix}$, and its determinant is $2 \cdot 8 - 2 \cdot 2 = 12$ again. This is not a coincidence: it is known that, in any number field $\mathbb{Q}(\gamma)$, the determinant⁴ of the integral trace form is the same for any choice of the integral basis. Hence, this determinant depends only on the number field itself, and it is called the *discriminant* of the number field $\mathbb{Q}(\gamma)$, see Sect. 5.9 for an alternative but equivalent definition of the discriminant. It is known that the discriminant of a number field is always an integer.

Counting Number Fields with Bounded Discriminant

Interestingly, there are no other number fields of degree 2 with discriminant 12. More generally, there are no two different (that is, non-isomorphic) number fields of degree 2 with the same discriminant. In other words, the discriminant of such a field can serve as a unique identifier for it. Unfortunately, this convenient property does not hold in general: for example, if γ_1 and γ_2 are solutions to the equations $x^3 - 21x + 28 = 0$ and $x^3 - 21x - 35 = 0$, respectively, then $\mathbb{Q}(\gamma_1)$ and $\mathbb{Q}(\gamma_2)$ are different number fields of degree 3 with the same discriminant 3969. An important open question is to estimate the number $N_n(X)$ of different number fields which have degree n and absolute value of the discriminant at most X . In fact, there is an old and difficult conjecture stating that, for every $n > 1$,

$$\lim_{X \rightarrow \infty} \frac{N_n(X)}{X} = c_n,$$

⁴For a general $n \times n$ matrix, the definition of the determinant is a bit complicated. Let S_n be the set of all possible permutations $\sigma = (\sigma(1), \dots, \sigma(n))$ of the set $(1, 2, \dots, n)$. For example, $(3, 1, 2)$ is a possible permutation of $(1, 2, 3)$. For general n , there are $n!$ possible permutations. Each permutation σ can be “implemented” by starting from $(1, 2, \dots, n)$ and exchanging adjacent elements, e.g. $(1, 2, 3) \rightarrow (1, 3, 2) \rightarrow (3, 1, 2)$. The sign of sigma is defined as $(-1)^n$, where n is the number of steps in this sequence. The determinant of a matrix with entries t_{ij} is defined as $\sum_{\sigma \in S_n} \text{sign}(\sigma) \prod_{i=1}^n t_{\sigma(i)i}$.

where c_n is a positive constant. In other words, the conjecture states that there are about c_n number fields of degree n per discriminant on average. The conjecture is easy for $n = 2$, the case $n = 3$ was resolved in 1971 by Davenport and Heilbronn, [107] while the case $n = 4$ was proved by Bhargava [51] in 2005, see Sect. 5.9. In 2010, Bhargava [53] resolved the next case, $n = 5$.

Theorem 10.11 *Let $N_5(X)$ be the number of number fields of degree 5 such that the absolute value of the discriminant is at most X . Then*

$$\lim_{X \rightarrow \infty} \frac{N_5(X)}{X} = c_5 \approx 0.149\dots$$

Reference

M. Bhargava, The density of discriminants of quintic rings and fields, *Annals of Mathematics* **172**-3, (2010), 1559–1591.

10.12 On the Negative Pell Equation

An Equation Which Helps to Approximate $\sqrt{2}$

In ancient times, people discovered that the length of the diagonal of the square with unit side, $\sqrt{2}$, is not exactly equal to any ratio $\frac{x}{y}$ of integers x and y . Motivated by this, they started to look for such a ratio which approximates $\sqrt{2}$ from below and above as closely as possible. If $\frac{x}{y}$ is an approximation from below, then $\frac{x}{y} < \sqrt{2}$, or $x < \sqrt{2}y$, or $x^2 < 2y^2$, or, equivalently, $x^2 - 2y^2 < 0$. To get as good an approximation as possible, it is natural to find integers x and y such that $x^2 - 2y^2$ is as close to 0 as possible. Because $x^2 - 2y^2$ is an integer, and the closest negative integer to 0 is -1 , it is natural to look for a pair of integers x and y such that

$$x^2 - 2y^2 = -1. \tag{10.25}$$

As one can easily see, one solution is $x = y = 1$. The ratio $\frac{1}{1} = 1$ does indeed “approximate” $\sqrt{2}$, and this is the best approximation with denominator 1, that is, the best approximation of $\sqrt{2}$ by an integer. To find a better approximation using fractions, we need to find other solutions of (10.25). For this, the following observation is useful: if the pair (x, y) is a solution of (10.25), then so is the pair $(3x + 4y, 2x + 3y)$. Indeed, substituting this into (10.25), we get

$$(3x + 4y)^2 - 2(2x + 3y)^2 = 9x^2 + 2 \cdot 3x \cdot 4y + 16y^2 - 2(4x^2 + 2 \cdot 2x \cdot 3y + 9y^2) = x^2 - 2y^2,$$

hence, if $x^2 - 2y^2 = -1$, then also $(3x + 4y)^2 - 2(2x + 3y)^2 = -1$. Because $x = y = 1$ is a solution of (10.25), this rule implies that $x = 3 \cdot 1 + 4 \cdot 1 = 7$ and $y =$

$2 \cdot 1 + 3 \cdot 1 = 5$ is a solution as well. This can also be checked directly: indeed $7^2 - 2 \cdot 5^2 = 49 - 50 = -1$. The corresponding ratio $\frac{7}{5} = 1.4$ is the best approximation of $\sqrt{2}$ with denominator at most 5.

Because $x = 7, y = 5$ is a solution, the pair $(3x + 4y, 2x + 3y) = (3 \cdot 7 + 4 \cdot 5, 2 \cdot 7 + 3 \cdot 5) = (41, 29)$ is a solution as well, and indeed, $41^2 - 2 \cdot 29^2 = 1681 - 1682 = -1$, which gives an even better approximation $\frac{41}{29} \approx 1.4138$ to $\sqrt{2} \approx 1.4142$ from below. In this way we can generate an infinite sequence (x_n, y_n) of solutions to (10.25) : $x_0 = y_0 = 1$ and

$$x_{n+1} = 3x_n + 4y_n, \quad y_{n+1} = 2x_n + 3y_n, \quad n = 0, 1, 2, 3, \dots$$

This way we get a sequence $\frac{x_n}{y_n}$ of increasingly better approximations of $\sqrt{2}$ from below. Similarly, if the ratio $\frac{x}{y}$ approximates $\sqrt{2}$ from above, then the inequality $\frac{x}{y} > \sqrt{2}$ implies that $x^2 - 2y^2 > 0$. Because $x^2 - 2y^2$ is an integer, and the smallest positive integer is 1, we obtain the equation $x^2 - 2y^2 = 1$, and can use its solutions to construct a sequence of increasingly better approximations of $\sqrt{2}$ from above.

Approximations of \sqrt{d} , and Negative/Positive Pell Equations

More generally, if d is any positive integer which is not a perfect square, then \sqrt{d} is not exactly equal to any ratio $\frac{x}{y}$ of integers x and y , and we again need to look for approximations from below and above. If $\frac{x}{y}$ is an approximation from below, then $\frac{x}{y} < \sqrt{d}$ is equivalent to $x^2 - dy^2 < 0$, and, to find the best approximations, it is natural to look for integer solutions to the equation

$$x^2 - dy^2 = -1, \tag{10.26}$$

which is known as the *negative Pell equation*. Similarly, the search for an approximation of \sqrt{d} from above leads to the equation $x^2 - dy^2 = 1$, which is the (positive) Pell equation.

It is interesting to note that the name “Pell equation” originates from the work of Euler, who used this name... by mistake! In fact, the English mathematician John Pell had no connection with the study of this equation. But the name “Pell equation” is by now well-established, and it is difficult to change it. Sometimes, people use the name “Pell–Fermat equation”. In fact, Fermat studied it, but he was not the first, it was studied by Indian mathematicians many years before him!

For $d = 8$, $\sqrt{d} = \sqrt{8} = 2\sqrt{2}$, hence any approximation $\frac{x}{y}$ of $\sqrt{2}$ immediately implies the approximation $\frac{2x}{y}$ of $\sqrt{8}$. More generally, if $d = k^2 \cdot m$ for some integers $k > 1, m$, then $\sqrt{d} = k\sqrt{m}$, and the problem of approximating \sqrt{d} reduces to the problem of approximating \sqrt{m} . Hence, it is sufficient to develop a good approximation method for \sqrt{d} only for *square-free* integers d , that is, those which are not divisible by k^2 for any integer $k > 1$. The first square-free integers are

1, 2, 3, 5, 6, 7, 10, 11, 13, 14, 15, ..., and so on. Below, we will study Eq. (10.26) for square-free integers $d \geq 2$.

One Solution Generates Infinitely Many

If (x_0, y_0) is any integer solution to (10.26), then it can be used to construct the whole infinite sequence (x_n, y_n) of solutions to (10.26) in a similar way as we did for $d = 2$. The sequence is given by

$$x_{n+1} = ux_n + dy_n, \quad y_{n+1} = vx_n + uy_n, \quad n = 0, 1, 2, 3, \dots \quad (10.27)$$

where (u, v) is a solution to the positive Pell equation, that is, integers such that $u^2 - dv^2 = 1$. Indeed, direct calculation shows that

$$x_{n+1}^2 - dy_{n+1}^2 = (ux_n + dy_n)^2 - d(vx_n + uy_n)^2 = (u^2 - dv^2)(x_n^2 - dy_n^2) = x_n^2 - dy_n^2,$$

hence, if $x_0^2 - dy_0^2 = -1$, then $x_1^2 - dy_1^2 = x_0^2 - dy_0^2 = -1$, hence $x_2^2 - dy_2^2 = x_1^2 - dy_1^2 = -1$, and so on. This way we get infinite sequence (x_n, y_n) of solutions to (10.26), and the corresponding fractions $\frac{x_n}{y_n}$ give increasingly better approximations of \sqrt{d} from below. Approximations from above can be constructed in a similar way.

A Question About Solution Existence

This argument has two problems. First, it assumes the existence of a solution (u, v) of the positive Pell equation $u^2 - dv^2 = 1$. Of course, this equation always has solutions $(u, v) = (1, 0)$ and $(u, v) = (-1, 0)$, which are called “trivial”, but these solutions cannot be used to generate an infinite sequence in (10.27). Fortunately, it is known that whenever the positive integer d is not a perfect square, the positive Pell equation always has a non-trivial solution. Moreover, for many centuries people have developed ways to actually *find* such a non-trivial solution (u, v) .

The second problem is that we need an initial integer solution (x_0, y_0) of the negative Pell equation (10.26) to start the iterations (10.27). We have proved that if we have one solution (x_0, y_0) to (10.26), then we can find infinitely many, but we did not prove that this first solution exists. Indeed, it turns out that there are some values of d such that Eq. (10.26) does not have integer solutions at all. For example, this happens for $d = 3$. Indeed, assume that there exist integers x, y such that $x^2 - 3y^2 = -1$. Every integer x can be written as either $x = 3k$, or $x = 3k + 1$, or $x = 3k + 2$ for some integer k . If $x = 3k$, then $x^2 - 3y^2 = -1$ reduces to $9k^2 - 3y^2 = -1$, or $3(3k^2 - y^2) = -1$. But -1 is not divisible by 3, a contradiction. Similarly, if $x = 3k + 1$, then $(3k + 1)^2 - 3y^2 = -1$ reduces to $9k^2 + 6k + 1 - 3y^2 = -1$, or $3(3k^2 + 2k - y^2) = -2$, again a contradiction. Finally, with $x = 3k + 2$, $(3k + 2)^2 - 3y^2 = -1$ implies that $3(3k^2 + 4k - y^2) = -5$, which is again impossible

1	2	5	10	13	17	26	29	34	37	41	53	58	61	65	73
74	82	85	89	97	101	106	109	113	122	130	137	145	146	149	157
170	173	178	181	185	193	194	197	202	205	218	221	226	229	233	241
257	265	269	274	277	281	290	293	298	305	313	314	317	337	346	349

Fig. 10.12 The first 64 elements of S . Integers d such that (10.26) has no integer solutions are marked black

because -5 is not divisible by 3. In fact, the same proof works to show that Eq. (10.26) has no integer solutions for $d = 6$, and, more generally, for any d divisible by 3. Moreover, a similar argument works for $d = 7$, and also for any d divisible by 7. More generally, one can show that (10.26) has no integer solutions for any d divisible by any number of the form $4k + 3$ for some integers k .

Let S be the set of square-free integers which has no divisors of the form $4k + 3$. The first integers belonging to S are 1, 2, 5, 10, 13, 17, 26, 29, 34, 37 The above discussion implies that if a square-free integer d does not belong to S , then (10.26) has no integer solutions. What about the case $d \in S$? As we saw above, solutions exist for $d = 2$. For $d = 5$, we have a solution $x = 2$, $y = 1$, which can be used to generate an infinite sequence of solutions. One may also find a solution for $d = 10, 13, 17, 26$, and 29, but it turns out that (10.26) has no integer solutions for $d = 34$, despite the fact that $34 \in S$. In fact, (10.26) has no integer solutions for $d = 34, 146, 178, 194, 205, 221, 305$, and so on, despite the fact that all these values are element of S , see Fig. 10.12.

For How Many Values of d does a Solution Exist?

In general, for which square-free d does Eq. (10.26) have a solution? Can we at least count *how many* such values of d are there among the first, say, million positive integers? More generally, if $f(N)$ is the number of square-free integers d less than N such that Eq. (10.26) has an integer solution, can we estimate $f(N)$ for large N ? The following theorem, proved in [155], provides such an estimate.

Theorem 10.12 *Let $f(N)$ be defined as above. We have*

$$(\alpha - e(N))c \frac{N}{\sqrt{\ln N}} \leq f(N) \leq \left(\frac{2}{3} + e(N)\right) c \frac{N}{\sqrt{\ln N}}, \quad (10.28)$$

where c and $\alpha \approx 0.419$ are universal constants, and $e(N)$ is a function such that $\lim_{N \rightarrow \infty} e(N) = 0$.

In fact, it is known that, for large N , the number of integers belonging to the set S is about $c \frac{N}{\sqrt{\ln N}}$, where c is the same constant. Hence, Theorem 10.12 implies that the percentage of values of $d \in S$ such that Eq. (10.26) has a solution is between 41.9% and 66.7%.

Note that Eq. (10.26) has a solution for 57 values of d out of the first 64 elements of S depicted in Fig. 10.12. This is more than 89%, while the upper bound of Theorem 10.12 is just 66.7%. This shows that we should be very careful while making guesses about asymptotic questions using computations with small numbers. One partial explanation is that the negative Pell equation is always solvable for prime numbers in S , and there are many prime numbers at the beginning of the sequence S .

The Connection to Continued Fractions

Theorem 10.12 is a result about the existence of a solution. To apply (10.27), we actually need to *find* an initial solution (x_0, y_0) if it exists. This can be done using the theory of continued fractions.

The decimal expansion of an irrational number like $\sqrt{3} = 1.73205\dots$ can be viewed as a series of increasingly better approximations of $\sqrt{3}$ by rational numbers: $\sqrt{3} \approx 1.7$, $\sqrt{3} \approx 1.73$, $\sqrt{3} \approx 1.732$, and so on. There is another way to approximate $\sqrt{3}$ by a sequence of rational numbers. First, let us write $\sqrt{3} = 1 + \frac{1}{x_1}$, where $x_1 = \frac{1}{\sqrt{3}-1} \approx 1.36$. Approximating x_1 by $[x_1] = 1$, where the symbol $[x]$ denotes the largest integer not exceeding x , we obtain $\sqrt{3} \approx 1 + \frac{1}{1} = 2$. Next, writing $x_1 = [x_1] + \frac{1}{x_2}$, where $x_2 = \frac{1}{x_1 - [x_1]} \approx 2.73$, and approximating x_2 by $[x_2] = 2$, we obtain $x_1 \approx 1 + \frac{1}{2}$, and $\sqrt{3} \approx 1 + \frac{1}{1+\frac{1}{2}}$. In general, for every irrational number α , define $x_0 = \alpha$ and, iteratively, $x_{n+1} = \frac{1}{x_n - [x_n]}$, $n \geq 1$. Then

$$\alpha = [x_0] + \cfrac{1}{[x_1] + \cfrac{1}{[x_2] + \cfrac{1}{[x_3] + \dots}}},$$

and the sequence $[x_0], [x_1], [x_2], \dots$ is called the continued fraction expansion of α (see also Sect. 4.1).

It turns out that the continued fraction expansion of $\sqrt{3}$ is 1, 1, 2, 1, 2, 1, 2, 1, 2, ..., that is, the pattern 1, 2 repeats itself infinitely often. In general, if the expansion consists of some initial block followed by a pattern a_1, \dots, a_m which repeats itself infinitely often, then it is called *periodic*, and the minimal length m of such a pattern is called the *period* of the expansion. It is known that, for all integers d which are not perfect squares, the continued fraction expansion of \sqrt{d} is periodic with some period $m(d)$. For example, the continued fraction expansion of $\sqrt{2}$ is 1, 2, 2, 2, 2, 2, 2, ..., hence $m(2) = 1$. The expansion of $\sqrt{29}$ is 5, 2, 1, 1, 2, 10, 2, 1, 1, 2, 10, ..., with repeating pattern 2, 1, 1, 2, 10 of length 5, hence $m(29) = 5$. The expansion of $\sqrt{34}$ is 5, 1, 4, 1, 10, 1, 4, 1, 10, ... with repeating pattern 1, 4, 1, 10 and $m(34) = 4$.

There is a deep connection between the solvability of the negative Pell equation and continued fraction expansions. Namely, the negative Pell equation (10.26) has an integer solution if and only if $m(d)$ is an odd integer! Moreover, the continued fraction expansion of \sqrt{d} can be used to actually *find* a solution to (10.26) if it exists. Namely, if the expansion of \sqrt{d} is $a_0, a_1, \dots, a_m, a_1, \dots, a_m, \dots$ with repeating pattern

a_1, \dots, a_m , then $a_0 + 1/(a_1 + 1/(\dots a_{m-2} + 1/a_{m-1})\dots)$ is some rational number, which can be written as an irreducible fraction $\frac{x}{y}$. Then $x^2 - dy^2 = (-1)^{m(d)}$. In particular, (x, y) is a solution to (10.26) if $m(d)$ is odd. This solution is called the *fundamental solution*.

For example, $m(34) = 4$ is even, hence the equation $x^2 - 34y^2 = -1$ has no integer solutions. On the other hand, $m(29) = 5$ is odd, hence an integer solution to $x^2 - 29y^2 = -1$ exists. To find a solution, we use the continued fraction expansion of $\sqrt{29}$, which is $5, 2, 1, 1, 2, 10, 2, 1, 1, 2, 10\dots$, and calculate $5 + 1/(2 + 1/(1 + 1/(1 + 1/2))) = 70/13$, hence $x = 70, y = 13$ is the fundamental solution. Indeed, $70^2 - 29 \cdot 13^2 = 4900 - 4901 = -1$. We can now apply (10.27) to generate infinitely many other solutions.

Further Progress, and Open Questions

In a later work, Fouvry and Klüners [156] used deeper algebraic methods to improve the lower bound in (10.28) and show that Eq. (10.26) has a solution for at least $5\alpha/4 \approx 52.4\%$ of values of $d \in S$. Together with the upper bound 66.7% from Theorem 10.12, this result is a significant step towards a resolution of a Stevenhagen's conjecture [361], which predicts that the correct percentage should be about 58.1%.

Another important question is how large the fundamental solution is (obtained by the continued fraction method). It is reasonably small in our example above with $d = 29$, but, for example, for $d = 409$, x and y in the fundamental solution are larger than ten billion. It is conjectured that, for most of the d , the fundamental solution has size about $e^{\sqrt{d}}$, which is astronomical compared to the size of d . This is a very deep conjecture which is currently out of reach.

Reference

É. Fouvry and J. Klüners, On the negative Pell equation, *Annals of Mathematics* **172**-3, (2010), 2035–2104.

10.13 The Norms of Random Band Matrices

Linear Transformations of Lines and Planes

The function $f(x) = 2x$ can be viewed as a transformation of the coordinate line which sends every point $x \in \mathbb{R}$ to the point $2x$. The image of any interval (a, b) of length $b - a$ is the interval $(2a, 2b)$ of length $2(b - a)$, hence we say that this transformation “stretches” the coordinate line by a factor of 2. Similarly, the function $f(x) = 0.5x$ “contracts” the line by a factor of 2.

A transformation of the coordinate *plane* can be described by a function $A : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ sending any pair of real numbers (x, y) to another pair (u, v) . Here, we

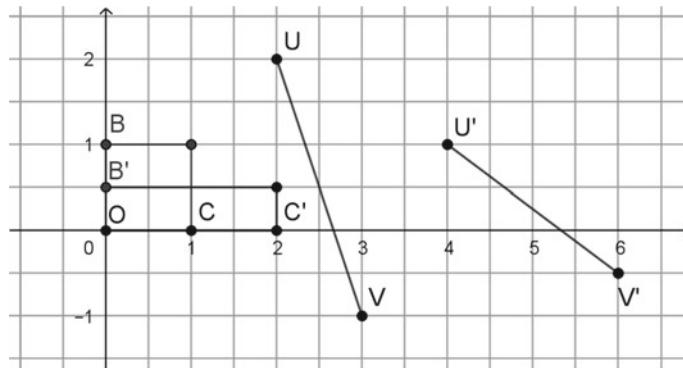


Fig. 10.13 Stretches and contractions in the transformation $A : (x, y) \rightarrow (2x, 0.5y)$

consider only linear transformations, sending the point $(0, 0)$ to itself. Such transformations can be described by a function of the form $(x, y) \rightarrow (ax + by, cx + dy)$ for some real coefficients a, b, c, d , e.g. $A : (x, y) \rightarrow (2x, 0.5y)$. Does this transformation “stretch” the plane or “contract” it? In fact, it “acts” differently on different regions of the plane. For example, if O, B, C are points in the plane with coordinates $(0, 0), (0, 1), (1, 0)$, respectively, then $|OB| = |OC| = 1$. Now, the transformation A sends point B to point B' with coordinates $(0, 0.5)$, hence it “contracts” the line segment OB by a factor of 2. On the other hand, A sends point C to point C' with coordinates $(2, 0)$, hence the line segment OC is stretched, see Fig. 10.13. See also Sect. 8.11 for more details and more examples of plane transformations.

The Maximal Stretching Factor

Does there exist a line segment which the transformation A stretches by a factor of more than 2? If $U = (x_1, y_1)$ and $V = (x_2, y_2)$ are arbitrary points, the length of the line interval UV is $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} = \sqrt{x^2 + y^2}$, where $x = x_2 - x_1$ and $y = y_2 - y_1$ are the coordinates of the vector \overrightarrow{UV} . If $U' = (2x_1, 0.5y_1)$ and $V' = (2x_2, 0.5y_2)$ are the images of U and V , respectively, then the length of $U'V'$ is $\sqrt{(2x_2 - 2x_1)^2 + (0.5y_2 - 0.5y_1)^2} = \sqrt{(2x)^2 + (0.5y)^2}$. To achieve maximal stretching, we need to maximize the value of the expression

$$\frac{\sqrt{(2x)^2 + (0.5y)^2}}{\sqrt{x^2 + y^2}}$$

With the new variable $u = \frac{x^2}{x^2 + y^2}$, this simplifies to $\sqrt{4u + 0.25(1-u)} = \sqrt{3.75u + 0.25}$. Because $0 \leq u \leq 1$, the maximum is attained if $u = 1$, and is equal to $\sqrt{3.75 \cdot 1 + 0.25} = 2$, hence, in this case, the stretching factor of 2 is the maximal possible.

As another example, consider the transformation of three-dimensional space sending any point (or vector) with coordinates (x, y, z) to the point (or vector) with coordinates $(y + z, x - z, x - y)$. What is the maximal stretching factor in this case? This corresponds to maximizing the expression

$$\frac{\sqrt{(y+z)^2 + (x-z)^2 + (x-y)^2}}{\sqrt{x^2 + y^2 + z^2}}.$$

In fact, $(y+z)^2 \leq (y+z)^2 + (y-z)^2 = 2(y^2 + z^2)$. Similarly, $(x-z)^2 \leq (x-z)^2 + (x+z)^2 = 2(x^2 + z^2)$ and $(x-y)^2 \leq 2(x^2 + y^2)$. Adding these together, we get

$$(y+z)^2 + (x-z)^2 + (x-y)^2 \leq 4(x^2 + y^2 + z^2),$$

which implies that $\frac{\sqrt{(y+z)^2 + (x-z)^2 + (x-y)^2}}{\sqrt{x^2 + y^2 + z^2}} \leq 2$. With $x = -1$, $y = z = 1$, equality holds, hence the maximal stretching is by a factor of 2.

***n*-variable Transformations, Matrices, and Their Norms**

In a similar way, one may study *n*-variable linear transformations of the form $(x_1, x_2, \dots, x_n) \rightarrow (y_1, y_2, \dots, y_n)$, where

$$y_j = a_{1j}x_1 + a_{2j}x_2 + \cdots + a_{nj}x_n, \quad j = 1, 2, \dots, n,$$

where a_{ij} , $i = 1, 2, \dots, n$, $j = 1, 2, \dots, n$, are coefficients. Usually, these coefficients are written as an $n \times n$ table (called a *matrix*), in which a_{ij} is written at the intersection of row i and column j . For example, the transformation $(x, y, z) \rightarrow (y + z, x - z, x - y)$, considered above, corresponds to the matrix

$$A^* = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & -1 \\ 1 & -1 & 0 \end{bmatrix}.$$

The *norm* of any $n \times n$ matrix A , denoted by $\|A\|$, is the maximal stretching factor of the corresponding transformation, that is, the maximal possible ratio

$$\frac{\sqrt{y_1^2 + y_2^2 + \cdots + y_n^2}}{\sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}}$$

over all choices of x_1, x_2, \dots, x_n , such that the denominator is non-zero. For example, $\|A^*\| = 2$.

The Norm of a “Typical” Matrix

The matrix A^* has some interesting properties. First, it is symmetric, that is, $a_{ij} = a_{ji}$ for all i and j . Second, all its entries are either 0, or 1, or -1 . Third, it has all 0’s on the main diagonal ($a_{ii} = 0$ for all i), but all entries which are “close” to the main diagonal but not on it are non-zero. Such matrices arise in applications, and an important question is to determine the norm of a “typical” matrix of this form. Here, by “typical” we mean that we select a matrix of this form at random, and would like to determine its norm with high probability.

Below is a formal model for random matrix selection. For every positive integer n , let A_n be a random $n \times n$ matrix whose elements a_{ij} are such that (i) $a_{ij} = a_{ji}$ for all i and j ; (ii) $a_{ij} = 0$ if either $i = j$ or $\min(|i - j|, n - |i - j|) > w_n$, where w_n is some constant, depending on n ; and (iii) all a_{ij} such that $0 < \min(|i - j|, n - |i - j|) \leq w_n$ are selected independently at random, each equal to 1 or -1 with equal chance. Random matrices A_n selected in accordance to these rules are known as *random band matrices*. For example, in case $n = 3$, $w_3 = 1$, the expression $\min(|i - j|, n - |i - j|)$ is equal to 0 and 1 if $i = j$ and $i \neq j$, respectively, and the conditions above reduce to (i) $a_{12} = a_{21}$, $a_{13} = a_{31}$, $a_{23} = a_{32}$; (ii) $a_{11} = a_{22} = a_{33} = 0$, and (iii) entries a_{12} , a_{13} , and a_{23} are selected independently at random, each equal to 1 or -1 with equal chance. For example, if a_{12} and a_{13} happened to be 1 and a_{23} happened to be -1 , we have got a matrix A^* whose norm is 2. In general, the norm $\|A_n\|$ of the randomly selected matrix A_n is a random number, which mathematicians call a *random variable*.

Convergence in Distribution

How can we predict or calculate something random? Sometimes it is possible! For example, if one tosses a fair coin with a 50%-50% chance of heads or tails, one cannot predict the result of a single experiment. However, if X_n denotes the fraction of heads in n experiments, then, for large n , one may confidently predict that X_n is “close” to 0.5. More formally, we say that a sequence X_n of random variables converges in distribution to a constant c if, for any $\varepsilon > 0$, $\lim_{n \rightarrow \infty} \mathbb{P}[c - \varepsilon < X_n < c + \varepsilon] = 1$, where \mathbb{P} denotes the probability.

In 2008, Khorunzhiy [223] conjectured that, if w_n grows faster than $\ln n$, then the norm of the matrix⁵ $A_n/(2\sqrt{2w_n})$ converges in distribution to 1. In 2010, this conjecture was proved by Sasha Sodin [350].

Theorem 10.13 *Let w_n be any sequence such that $w_n/\ln n \rightarrow +\infty$ (in other words, for any C there exists an N such that $w_n > C \ln n$ for all $n \geq N$). Then $\|A_n/2\sqrt{2w_n}\|$ converges in distribution to 1.*

⁵By the ratio of the matrix A_n and the constant $c = 2\sqrt{2w_n}$ we mean the matrix whose entries are the entries of A_n each divided by c .

Theorem 10.13 implies that, for large n , $\|A_n/2\sqrt{2w_n}\|$ is very close to 1 with high probability. This implies that $\|A_n\|$ is, up to some small relative error, approximately equal to $2\sqrt{2w_n}$.

It is known [61] that condition $w_n/\ln n \rightarrow +\infty$ cannot be removed, and, moreover, it is “sharp” in the sense that the conclusion of the theorem does not hold if $w_n/\ln n \rightarrow 0$.

In fact, Sodin developed a deep theory investigating various properties of the random matrices A_n arising from the model described above. Theorem 10.13 is just one of many corollaries of this theory.

Reference

S. Sodin, The spectral edge of some random band matrices, *Annals of Mathematics* **172**-3, (2010), 2223–2251.

Chapter 11

Further Reading



The following table summarizes the references which have been used during the preparation of each theorem's description, as well as some suggestions for further reading on the corresponding topic.

Thm.	References	Thm.	References
1.1	[64, 119, 210, 355, 363, 384]	3.6	[47, 89, 261]
1.2	[172, 187, 240]	3.7	[90, 218, 263, 340]
1.3	[258, 336, 376]	3.8	[68, 109, 328, 336]
1.4	[279, 281, 296, 299]	4.1	[277, 301, 302]
1.5	[77, 198, 199]	4.2	[307, 342, 351, 352]
1.6	[14, 28, 242]	4.3	[358, 381, 383, 405]
1.7	[209, 215, 235, 236, 273, 274, 393]	4.4	[88, 210, 384, 385]
2.1	[141, 142, 163]	4.5	[84, 169, 170, 241, 374]
2.2	[16, 54, 55, 58, 232]	4.6	[17, 18, 305]
2.3	[117, 204, 318, 319]	4.7	[48–50, 52, 114]
2.4	[85, 262, 346, 382]	4.8	[38, 116, 362]
2.5	[153, 190, 191, 207, 282, 317, 341]	4.9	[8, 39, 40, 300]
2.6	[193, 306, 406]	4.10	[5, 105, 276, 278, 311, 371]
2.7	[66, 112, 256]	5.1	[157, 206, 397]
2.8	[56, 278, 332]	5.2	[144, 149, 360, 402]
2.9	[56, 99, 255, 290, 332]	5.3	[98, 267, 268]
2.10	[115, 197, 284]	5.4	[25, 66, 280, 297]
2.11	[34, 167, 379]	5.5	[36, 60, 73, 138, 174, 175, 326, 331]
2.12	[37, 233, 234, 292, 315, 316, 323, 407]	5.6	[181, 211, 380]
2.13	[75, 162, 180, 227]	5.7	[27, 101, 118, 192, 224, 226]
3.1	[194, 244, 245, 399]	5.8	[35, 270, 272]
3.2	[106, 135, 237]	5.9	[51, 53, 107]
3.3	[46, 148, 269]	5.10	[184–186]
3.4	[86, 94, 96, 392]	5.11	[1, 2, 9, 63, 136, 254]
3.5	[87, 139, 324]	6.1	[133, 321, 359, 366]

(continued)

(Continued)

Thm.	References	Thm.	References
6.2	[93, 126, 335]	8.13	[62, 257, 368, 394]
6.3	[81, 82, 248, 249, 303, 395]	9.1	[111, 113, 333]
6.4	[44, 91, 92, 230, 251, 252]	9.2	[145–147, 149]
6.5	[125, 247, 336, 390]	9.3	[267, 275, 398]
6.6	[71, 330, 367]	9.4	[213, 329, 370, 372, 377]
6.7	[41, 43, 120, 222]	9.5	[19, 20, 289, 325]
7.1	[103, 250, 312]	9.6	[24, 26, 238]
7.2	[3, 4, 150]	9.7	[10, 11, 401]
7.3	[173, 259, 327]	9.8	[67, 308, 396]
7.4	[23, 389, 391]	9.9	[65, 127, 164, 165, 171, 309, 408]
7.5	[121, 122, 205]	9.10	[285, 310, 353]
7.6	[112, 231, 298, 348]	9.11	[7, 100, 137, 143, 212, 314, 375]
7.7	[69, 239, 334, 349]	9.12	[182, 189, 201, 266, 313, 322, 354, 378]
7.8	[31, 42, 221, 336]	9.13	[59, 95, 96, 228, 229, 304]
7.9	[32, 79, 108, 388]	9.14	[202, 217, 246, 343, 345, 403]
7.10	[158, 166, 168, 365]	10.1	[15, 214, 225, 283, 404]
8.1	[83, 97, 267]	10.2	[12, 21, 22, 78, 123, 293, 294]
8.2	[175, 176, 373, 386]	10.3	[179, 183, 337, 338]
8.3	[195, 216, 344]	10.4	[208, 347, 387]
8.4	[29, 196, 286]	10.5	[33, 151, 243, 356]
8.5	[74, 204, 253, 286]	10.6	[57, 72, 80, 369]
8.6	[124, 210, 219, 220, 357]	10.7	[13, 161, 265]
8.7	[70, 128, 247]	10.8	[175–178]
8.8	[30, 104, 287, 291, 304, 339]	10.9	[45, 76, 102, 203]
8.9	[6, 264, 271]	10.10	[200, 288, 295]
8.10	[131, 132, 134, 154]	10.11	[51, 53, 107, 208]
8.11	[110, 329, 364, 370]	10.12	[155, 156, 361, 400]
8.12	[152, 159, 188]	10.13	[61, 129, 130, 160, 223, 350]

References

1. Achlioptas, D., Naor, A.: The two possible values of the chromatic number of a random graph. *Ann. Math.* **162**(3), 1335–1351 (2005)
2. Achlioptas, D., Naor, A., Peres, Y.: Rigorous location of phase transitions in hard optimization problems. *Nature* **435**(7043), 759 (2005)
3. Adamczewski, B., Bugeaud, Y.: On the complexity of algebraic numbers. II. Continued fractions. *Acta Math.* **195**(1), 1–20 (2005)
4. Adamczewski, B., Bugeaud, Y.: On the complexity of algebraic numbers, I. Expansions in integer bases. *Ann. Math.* **165**(2), 547–565 (2007)
5. Agrawal, M., Kayal, N., Saxena, N.: PRIMES is in P. *Ann. Math.* **160**(2), 781–793 (2004)
6. Aharoni, I., Maurey, B., Mityagin, B.S.: Uniform embeddings of metric spaces and of Banach spaces into Hilbert spaces. *Isr. J. Math.* **52**(3), 251–265 (1985)
7. Aigner, M., Ziegler, G.M., Hofmann, K.H., Erdős, P.: *Proofs from the Book*, vol. 274. Springer, Berlin (2010)
8. Aldous, D.: The continuum random tree. I. *Ann. Probab.* **19**(1), 1–28 (1991)
9. Alon, N., Krivelevich, M.: The concentration of the chromatic number of random graphs. *Combinatorica* **17**(3), 303–313 (1997)
10. Alon, N., Shapira, A.: Every monotone graph property is testable. *SIAM J. Comput.* **38**(2), 505–522 (2008)
11. Alon, N., Shapira, A., Sudakov, B.: Additive approximation for edge-deletion problems. *Ann. Math.* **170**(1), 371–411 (2009)
12. Andrews, G.E., Berndt, B.C.: *Ramanujan’s Lost Notebook*, vol. 1. Springer, Berlin (2005)
13. Apostol, T.M.: *Introduction to Analytic Number Theory*. Springer Science & Business Media, Berlin (2013)
14. Arad, Z., Herzog, M.: *Products of Conjugacy Classes in Groups*, vol. 1112. Springer, Berlin (2006)
15. Arrow, K.J.: A difficulty in the concept of social welfare. *J. Polit. Econ.* **58**(4), 328–346 (1950)
16. Artin, E.: Theorie der zöpfe. In: *Abhandlungen aus dem mathematischen Seminar der Universität Hamburg*, vol. 4, pp. 47–72. Springer (1925)
17. Artstein, S., Milman, V., Szarek, S.J.: Duality of metric entropy. *Ann. Math.* **159**(3), 1313–1328 (2004)
18. Artstein-Avidan, S., Giannopoulos, A., Milman, V.D.: *Asymptotic Geometric Analysis*, Part I, vol. 202. American Mathematical Society, Providence (2015)
19. Artstein-Avidan, S., Milman, V.: A characterization of the concept of duality. *Electron. Res. Announc. Math. Sci.* **14**, 42–59 (2007)

20. Artstein-Avidan, S., Milman, V.: The concept of duality in convex analysis, and the characterization of the Legendre transform. *Ann. Math.* **169**(2), 661–674 (2009)
21. Atkin, A., O'Brien, J.: Some properties of $p(n)$ and $c(n)$ modulo powers of 13. *Trans. Am. Math. Soc.* **126**(3), 442–459 (1967)
22. Atkin, A.O.L., Swinnerton-Dyer, P.: Some properties of partitions. *Proc. Lond. Math. Soc.* **3**(1), 84–106 (1954)
23. Avila, A., Forni, G.: Weak mixing for interval exchange transformations and translation flows. *Ann. Math.* **165**(2), 637–664 (2007)
24. Avila, A., Jitomirskaya, S.: The ten martini problem. *Ann. Math.* **170**(1), 303–342 (2009)
25. Avila, A., Moreira, C.G.: Statistical properties of unimodal maps: the quadratic family. *Ann. Math.* **161**(2), 831–881 (2005)
26. Azbel, M.Y.: Energy spectrum of a conduction electron in a magnetic field. *Sov. Phys. JETP* **19**(3), 634–645 (1964)
27. Babai, L., Fortnow, L., Lund, C.: Non-deterministic exponential time has two-prover interactive protocols. *Comput. Complex.* **1**(1), 3–40 (1991)
28. Babai, L., Hetyei, G., Kantor, W.M., Lubotzky, A., Seress, A.: On the diameter of finite groups. In: 1990 Proceedings of 31st Annual Symposium on Foundations of Computer Science, pp. 857–865. IEEE (1990)
29. Babai, L., Seress, Á.: On the diameter of permutation groups. *Eur. J. Comb.* **13**(4), 231–243 (1992)
30. Bachoc, C., Vallentin, F.: New upper bounds for kissing numbers from semidefinite programming. *J. Am. Math. Soc.* **21**(3), 909–924 (2008)
31. Baker, R.C.: Dirichlet's theorem on Diophantine approximation. *Math. Proc. Camb. Philos. Soc.* **83**, 37–59 (1978)
32. Baker, R.C.: Diagonal cubic equations. II. *Acta Arith.* **53**(3), 217–250 (1989)
33. Balázs, M., Seppäläinen, T.: Order of current variance and diffusivity in the asymmetric simple exclusion process. *Ann. Math.* **171**(2), 1237–1265 (2010)
34. Banach, S.: Théorie des opérations linéaires. *Monografje Matematyczne* **1**, (1932)
35. Bartal, Y., Linial, N., Mendel, M., Naor, A.: On metric Ramsey-type phenomena. *Ann. Math.* **162**(2), 643–709 (2005)
36. Behrend, F.A.: On sets of integers which contain no three terms in arithmetical progression. *Proc. Natl. Acad. Sci. USA* **32**(12), 331 (1946)
37. Behrend, R.E., Fischer, I., Konvalinka, M.: Diagonally and anti-diagonally symmetric alternating sign matrices of odd order. *Adv. Math.* **315**, 324–365 (2017)
38. Belius, D., Kistler, N.: The subleading order of two dimensional cover times. *Probab. Theory Relat. Fields* **167**(1–2), 461–552 (2017)
39. Benjamini, I., Kesten, H., Peres, Y., Schramm, O.: Geometry of the uniform spanning forest: transitions in dimensions 4, 8, 12. *Ann. Math.* **160**(2), 465–491 (2004)
40. Benjamini, I., Lyons, R., Peres, Y., Schramm, O.: Special invited paper: uniform spanning forests. *Ann. Probab.* **29**(1), 1–65 (2001)
41. Beresnevich, V., Dickinson, D., Velani, S.: Measure theoretic laws for lim sup sets. *Mem. Am. Math. Soc.* **846**, 91 (2006)
42. Beresnevich, V., Dickinson, D., Velani, S.: Diophantine approximation on planar curves and the distribution of rational points. *Ann. Math.* **166**(2), 367–426 (2007)
43. Beresnevich, V., Velani, S.: A mass transference principle and the Duffin-Schaeffer conjecture for Hausdorff measures. *Ann. Math.* **164**(3), 971–992 (2006)
44. Berge, C.: Farbung von Graphen, deren sämtliche bzw. deren ungerade Kreise starr sind. *Wissenschaftliche Zeitschrift* (1961)
45. Berger, R.: The Undecidability of the Domino Problem, vol. 66. American Mathematical Society, Providence (1966)
46. Bernal, A.: A note on the one-dimensional maximal function. *Proc. R. Soc. Edinb. Sect. A Math.* **111**(3–4), 325–328 (1989)
47. Berry, M.V., Tabor, M.: Level clustering in the regular spectrum. *Proc. R. Soc. Lond. A* **356**(1686), 375–394 (1977)

48. Bhargava, M.: Higher composition laws I: A new view on Gauss composition, and quadratic generalizations. *Ann. Math.* **159**(1), 217–250 (2004)
49. Bhargava, M.: Higher composition laws II: On cubic analogues of Gauss composition. *Ann. Math.* **159**(2), 865–886 (2004)
50. Bhargava, M.: Higher composition laws III: The parametrization of quartic rings. *Ann. Math.* **159**(3), 1329–1360 (2004)
51. Bhargava, M.: The density of discriminants of quartic rings and fields. *Ann. Math.* **162**(2), 1031–1063 (2005)
52. Bhargava, M.: Higher composition laws IV: The parametrization of quintic rings. *Ann. Math.* **167**(1), 53–94 (2008)
53. Bhargava, M.: The density of discriminants of quintic rings and fields. *Ann. Math.* **172**(3), 1559–1591 (2010)
54. Bigelow, S.: The Burau representation is not faithful for $n = 5$. *Geom. Topol.* **3**(1), 397–404 (1999)
55. Bigelow, S.: Braid groups are linear. *J. Am. Math. Soc.* **14**(2), 471–486 (2001)
56. Birget, J.C., Ol'Shanskii, A.Y., Rips, E., Sapir, M.V.: Isoperimetric functions of groups and computational complexity of the word problem. *Ann. Math.* **156**(2), 467–518 (2002)
57. Birkhoff, G.D.: Proof of the ergodic theorem. *Proc. Natl. Acad. Sci.* **17**(12), 656–660 (1931)
58. Birman, J.: Review of Braid groups are linear groups by S. Bachmuth, MR 98h **20061** (1998)
59. Blichfeldt, H.F.: The minimum values of positive quadratic forms in six, seven and eight variables. *Mathematische Zeitschrift* **39**(1), 1–15 (1935)
60. Bloom, T.F.: A quantitative improvement for Roth's theorem on arithmetic progressions. *J. Lond. Math. Soc.* **93**(3), 643–663 (2016)
61. Bogachev, L.V., Molchanov, S.A., Pastur, L.A.: On the level density of random band matrices. *Math. Notes* **50**(6), 1232–1242 (1991)
62. Bogachev, V.I.: Measure Theory, vol. 1. Springer Science & Business Media, Berlin (2007)
63. Bollobás, B.: Random graphs. *Modern Graph Theory*, pp. 215–252. Springer, Berlin (1998)
64. Bolthausen, E.: On the volume of the Wiener sausage. *Ann. Probab.* **18**(4), 1576–1582 (1990)
65. Bombieri, E., Friedlander, J.B., Iwaniec, H.: Primes in arithmetic progressions to large moduli. *Acta Math.* **156**(1), 203–251 (1986)
66. Bonatti, C., Díaz, L.J., Viana, M.: Dynamics Beyond Uniform Hyperbolicity: A Global Geometric and Probabilistic Perspective, vol. 102. Springer Science & Business Media, Berlin (2006)
67. Borcea, J., Brändén, P.: Pólya-Schur master theorems for circular domains and their boundaries. *Ann. Math.* **170**(1), 465–492 (2009)
68. Borel, E.: Contribution à l'analyse arithmétique du continu. *Journal de mathématiques pures et appliquées* **9**, 329–375 (1903)
69. Borwein, P., Dobrowolski, E., Mossinghoff, M.J.: Lehmer's problem for polynomials with odd coefficients. *Ann. Math.* **166**(2), 347–366 (2007)
70. Borwein, P., Erdélyi, T., Ferguson, R., Lockhart, R.: On the zeros of cosine polynomials: solution to a problem of Littlewood. *Ann. Math.* **167**(3), 1109–1117 (2008)
71. Boucheron, S., Bousquet, O., Lugosi, G.: Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.* **9**, 323–375 (2005)
72. Bourgain, J.: Pointwise ergodic theorems for arithmetic sets. *Publications Mathématiques de l'Institut des Hautes Études Scientifiques* **69**(1), 5–41 (1989)
73. Bourgain, J.: On triples in arithmetic progression. *Geom. Funct. Anal.* **9**(5), 968–984 (1999)
74. Bourgain, J., Gamburd, A.: Uniform expansion bounds for Cayley graphs of $SL_2(F_p)$. *Ann. Math.* **167**(2), 625–642 (2008)
75. Bourgain, J., Lindenstrauss, J., Milman, V.: Minkowski sums and symmetrizations. *Geometric Aspects of Functional Analysis*, pp. 44–66. Springer, Berlin (1988)
76. Boyle, M.: Open problems in symbolic dynamics. *Contemp. Math.* **469**, 69–118 (2008)
77. Brauer, R.: Representations of finite groups. *Lect. Mod. Math.* **1**, 133–175 (1963)
78. Bringmann, K., Ono, K.: Dyson's ranks and Maass forms. *Ann. Math.* **171**(1), 419–449 (2010)

79. Brüdern, J., Wooley, T.D.: The Hasse principle for pairs of diagonal cubic forms. *Ann. Math.* **166**(3), 865–895 (2007)
80. Buczolich, Z., Mauldin, R.D.: Divergent square averages. *Ann. Math.* **171**(3), 1479–1530 (2010)
81. Bugeaud, Y., Mignotte, M., Siksek, S.: Classical and modular approaches to exponential Diophantine equations I. Fibonacci and Lucas perfect powers. *Ann. Math.* **163**(3), 969–1018 (2006)
82. Bugeaud, Y., Mignotte, M., Siksek, S.: Classical and modular approaches to exponential Diophantine equations II. The Lebesgue-Nagell equation. *Compos. Math.* **142**(1), 31–62 (2006)
83. Calabi, E.: Problems in differential geometry. In: Proceedings of the United States-Japan Seminar in Differential Geometry, Kyoto, Japan, p. 170 (1965)
84. Carleson, L.: On convergence and growth of partial sums of Fourier series. *Acta Math.* **116**(1), 135–157 (1966)
85. Carlson, J.A., Jaffe, A., Wiles, A.: The Millennium Prize Problems. American Mathematical Society, Providence (2006)
86. Chang, H.C., Wang, L.C.: A simple proof of Thue’s Theorem on circle packing (2010). arXiv preprint [arXiv:1009.4322](https://arxiv.org/abs/1009.4322)
87. Chang, M.C.: The Erdős-Szemerédi problem on sum set and product set. *Ann. Math.* **157**(3), 939–957 (2003)
88. Chen, X.: Random Walk Intersections: Large Deviations and Related Topics. Mathematical Surveys and Monographs, vol. 157. American Mathematical Society, Providence (2010)
89. Cheng, Z., Lebowitz, J.L.: Statistics of energy levels in integrable quantum systems. *Phys. Rev. A* **44**(6), R3399 (1991)
90. Cheung, Y.: Hausdorff dimension of the set of nonergodic directions. *Ann. Math.* **158**(2), 661–678 (2003)
91. Chudnovsky, M., Cornuéjols, G., Liu, X., Seymour, P., Vušković, K.: Recognizing berge graphs. *Combinatorica* **25**(2), 143–186 (2005)
92. Chudnovsky, M., Robertson, N., Seymour, P., Thomas, R.: The strong perfect graph theorem. *Ann. Math.* **164**(1), 51–229 (2006)
93. Cohen, H.: Constructing and counting number fields. In: Proceedings of International Congress of Mathematicians (Beijing 2002), pp. 129–138 (2002)
94. Cohn, H., Elkies, N.: New upper bounds on sphere packings I. *Ann. Math.* **157**(2), 689–714 (2003)
95. Cohn, H., Kumar, A.: Optimality and uniqueness of the Leech lattice among lattices. *Ann. Math.* **170**(3), 1003–1050 (2009)
96. Cohn, H., Kumar, A., Miller, S.D., Radchenko, D., Viazovska, M.: The sphere packing problem in dimension 24. *Ann. Math.* **185**(3), 1017–1033 (2017)
97. Colding, T.H., Minicozzi, W.P.: The Calabi-Yau conjectures for embedded surfaces. *Ann. Math.* **167**(1), 211–243 (2008)
98. Colding, T.H., Minicozzi, W.P., et al.: Complete properly embedded minimal surfaces in R^3 . *Duke Math. J.* **107**(2), 421–426 (2001)
99. Collins, D.J., et al.: A simple presentation of a group with unsolvable word problem. III. *J. Math.* **30**(2), 230–234 (1986)
100. Conlon, D.: A new upper bound for diagonal Ramsey numbers. *Ann. Math.* **170**(2), 941–960 (2009)
101. Cook, S.A.: The complexity of theorem-proving procedures. In: Proceedings of the Third Annual ACM Symposium on Theory of Computing, pp. 151–158. ACM (1971)
102. Cooper, S.B.: Computability Theory. Chapman and Hall/CRC, Boca Raton (2017)
103. Coppersmith, D., Rivlin, T.J.: The growth of polynomials bounded at equally spaced points. *SIAM J. Math. Anal.* **23**(4), 970–983 (1992)
104. Coxeter, H.S.: An upper bound for the number of equal nonoverlapping spheres that can touch another of the same size. In: Convexity: Proceedings of the Seventh Symposium in Pure Mathematics of the American Mathematical Society, vol. 7, pp. 53–71. American Mathematical Society (1963)

105. Crandall, R., Pomerance, C.B.: Prime Numbers: A Computational Perspective, vol. 182. Springer Science & Business Media, Berlin (2006)
106. Croft III, E.S.: On a coloring conjecture about unit fractions. *Ann. Math.* **157**(2), 545–556 (2003)
107. Davenport, H., Heilbronn, H.: On the density of discriminants of cubic fields. II. *Proc. R. Soc. Lond. Ser. A Math. Phys. Sci.* **322**, 405–420 (1971)
108. Davenport, H., Lewis, D.J.: Simultaneous equations of additive type. *Phil. Trans. R. Soc. Lond. A* **264**(1155), 557–595 (1969)
109. Davenport, H., Schmidt, W.: Approximation to real numbers by algebraic integers. *Acta Arith.* **15**(4), 393–416 (1969)
110. Davidson, K.R., Szarek, S.J.: Local operator theory, random matrices and Banach spaces. *Handbook of the Geometry of Banach Spaces* **1**(317–366), 131 (2001)
111. De Giorgi, E.: Convergence problems for functionals and operators. In: *Proceedings of International Meeting on Recent Methods in Nonlinear Analysis*, pp. 131–188 (1978)
112. De Melo, W., Van Strien, S.: One-dimensional dynamics, vol. 25. Springer Science & Business Media, Berlin (2012)
113. Del Pino, M., Kowalczyk, M., Wei, J.: On De Giorgi’s conjecture in dimension $N \geq 9$. *Ann. Math.* **174**(3), 1485–1569 (2011)
114. Delone, B.N., Faddeev, D.: *The Theory of Irrationalities of the Third Degree*. American Mathematical Society, Providence (1964)
115. Delzell, C.: A continuous, constructive solution to Hilbert’s 17-th problem. *Inventiones mathematicae* **76**(3), 365–384 (1984)
116. Dembo, A., Peres, Y., Rosen, J., Zeitouni, O.: Cover times for Brownian motion and random walks in two dimensions. *Ann. Math.* **160**(2), 433–464 (2004)
117. Dinur, I.: The PCP theorem by gap amplification. *J. ACM (JACM)* **54**(3), 12 (2007)
118. Dinur, I., Safra, S.: On the hardness of approximating minimum vertex cover. *Ann. Math.* **162**(1), 439–485 (2005)
119. Donsker, M., Varadhan, S.: Asymptotics for the Wiener sausage. *Commun. Pure Appl. Math.* **28**(4), 525–565 (1975)
120. Duffin, R.J., Schaeffer, A.C., et al.: Khintchines problem in metric Diophantine approximation. *Duke Math. J.* **8**(2), 243–255 (1941)
121. Dugger, D., Isaksen, D.C.: Algebraic K-theory and sums-of-squares formulas. *Documenta Mathematica* **10**, 357–366 (2005)
122. Dugger, D., Isaksen, D.C.: The Hopf condition for bilinear forms over arbitrary fields. *Ann. Math.* **165**(3), 943–964 (2007)
123. Dyson, F.J.: Some guesses in the theory of partitions. *Eureka (Cambridge)* **8**(10), 10–15 (1944)
124. Eden, M.: A two-dimensional growth process. *Dynamics of Fractal Surfaces*, vol. 4, pp. 223–239. World Scientific, London (1961)
125. Einsiedler, M., Katok, A., Lindenstrauss, E.: Invariant measures and the set of exceptions to Littlewood’s conjecture. *Ann. Math.* **164**(2), 513–560 (2006)
126. Ellenberg, J.S., Venkatesh, A.: The number of extensions of a number field with fixed degree and bounded discriminant. *Ann. Math.* **163**(2), 723–741 (2006)
127. Elliott, P., Halberstam, H.: A conjecture in prime number theory. *Symp. Math.* **4**, 59–72 (1968)
128. Erdélyi, T.: The number of unimodular zeros of self-reciprocal polynomials with coefficients in a finite set. *Acta Arith.* **176**(2), 177–200 (2016)
129. Erdős, L.: Universality of Wigner random matrices: a survey of recent results. *Russ. Math. Surv.* **66**(3), 507 (2011)
130. Erdős, L., Yau, H.T.: Universality of local spectral statistics of random matrices. *Bull. Am. Math. Soc.* **49**(3), 377–414 (2012)
131. Erdős, P.: Note on sequences of integers no one of which is divisible by any other. *J. Lond. Math. Soc.* **1**(2), 126–128 (1935)
132. Erdős, P.: An asymptotic inequality in the theory of numbers. *Vestnik Leningrad Univ* **15**, 41–49 (1960)

133. Erdős, P.: On the representation of large integers as sums of distinct summands taken from a fixed set. *Acta. Arith.* **7**, 345–354 (1962)
134. Erdős, P.: Some remarks on number theory. *Isr. J. Math.* **3**(1), 6–12 (1965)
135. Erdős, P., Graham, R.L.: Old and new problems and results in combinatorial number theory, vol. 28. L’Enseignement mathématique (1980)
136. Erdős, P., Rényi, A.: On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* **5**(1), 17–60 (1960)
137. Erdős, P., Szekeres, G.: A combinatorial problem in geometry. *Compositio mathematica* **2**, 463–470 (1935)
138. Erdős, P., Turán, P.: On some sequences of integers. *J. Lond. Math. Soc.* **1**(4), 261–264 (1936)
139. Erdős, P., Szemerédi, E.: On sums and products of integers. *Studies in Pure Mathematics*, pp. 213–218. Springer, Berlin (1983)
140. Erdős, P., et al.: The difference of consecutive primes. *Duke Math. J.* **6**(2), 438–441 (1940)
141. Eremenko, A., Gabrielov, A.: Rational functions with real critical points and the B. and M. Shapiro conjecture in real enumerative geometry. *Ann. Math.* **155**(1), 105–129 (2002)
142. Eremenko, A., Gabrielov, A.: An elementary proof of the B. and M. Shapiro conjecture for rational functions. *Notions of Positivity and the Geometry of Polynomials*, pp. 167–178. Springer, Berlin (2011)
143. Evans, R.J., Pulham, J.R., Sheehan, J.: On the number of complete subgraphs contained in certain graphs. *J. Comb. Theory Ser. B* **30**(3), 364–371 (1981)
144. Fefferman, C.: A sharp form of Whitney’s extension theorem. *Ann. Math.* **161**(1), 509–577 (2005)
145. Fefferman, C.: Fitting a C^m -smooth function to data. III. *Ann. Math.* **170**(1), 427–441 (2009)
146. Fefferman, C., Klartag, B.: Fitting a C^m -smooth function to data. I. *Ann. Math.* **169**(1), 315–346 (2009)
147. Fefferman, C., Klartag, B., et al.: Fitting a C^m -smooth function to data II. *Revista Matemática Iberoamericana* **25**(1), 49–273 (2009)
148. Fefferman, C., Stein, E.M.: Some maximal inequalities. *Am. J. Math.* **93**(1), 107–115 (1971)
149. Fefferman, C.L.: Whitneys extension problems and interpolation of data. *Bull. Am. Math. Soc.* **46**(2), 207–220 (2009)
150. Ferenczi, S., Mauduit, C.: Transcendence of numbers with a low complexity expansion. *J. Number Theory* **67**(2), 146–161 (1997)
151. Ferrari, P.A., Fontes, L.R.: Current fluctuations for the asymmetric simple exclusion process. *Ann. Probab.* **22**(2), 820–832 (1994)
152. Figalli, A., Maggi, F., Pratelli, A.: A mass transportation approach to quantitative isoperimetric inequalities. *Inventiones mathematicae* **182**(1), 167–211 (2010)
153. Foisy, J., Alfaro Garcia, M., Brock, J., Hodges, N., Zimba, J.: The standard double soap bubble in \mathbb{R}^2 uniquely minimizes perimeter. *Pac. J. Math.* **159**(1), 47–59 (1993)
154. Ford, K.: The distribution of integers with a divisor in a given interval. *Ann. Math.* **168**(2), 367–433 (2008)
155. Fouvry, É., Klüners, J.: On the negative Pell equation. *Ann. Math.* **172**(3), 2035–2104 (2010)
156. Fouvry, É., Klüners, J.: The parity of the period of the continued fraction of \sqrt{d} . *Proc. Lond. Math. Soc.* **101**(2), 337–391 (2010)
157. Furstenberg, H.: Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. *Journal d’Analyse Mathématique* **31**(1), 204–256 (1977)
158. Furstenberg, H., Katznelson, Y.: An ergodic Szemerédi theorem for commuting transformations. *Journal d’Analyse Mathématique* **34**(1), 275–291 (1978)
159. Fusco, N., Maggi, F., Pratelli, A.: The sharp quantitative isoperimetric inequality. *Ann. Math.* **168**(3), 941–980 (2008)
160. Fyodorov, Y.V., Mirlin, A.D.: Statistical properties of eigenfunctions of random quasi 1d one-particle Hamiltonians. *Int. J. Mod. Phys. B* **8**(27), 3795–3842 (1994)
161. Gelfond, A.: Sur les nombres qui ont des propriétés additives et multiplicatives données. *Acta Arith.* **13**(3), 259–265 (1968)

162. Giannopoulos, A., Milman, V.: Asymptotic convex geometry short overview. Different Faces of Geometry, pp. 87–162. Springer, Berlin (2004)
163. Goldberg, L.R.: Catalan numbers and branched coverings by the Riemann sphere. *Adv. Math.* **85**(2), 129–144 (1991)
164. Goldston, D.A., Pintz, J., Yalçın Yıldırım, C.: Primes in tuples II. *Acta Math.* **204**(1), 1–47 (2010)
165. Goldston, D.A., Pintz, J., Yıldırım, C.Y.: Primes in tuples I. *Ann. Math.* **170**(2), 819–862 (2009)
166. Gowers, W.T.: A new proof of Szemerédi’s theorem. *Geom. Funct. Anal. GAFA* **11**(3), 465–588 (2001)
167. Gowers, W.T.: An infinite Ramsey theorem and some Banach-space dichotomies. *Ann. Math.* **156**(3), 797–833 (2002)
168. Gowers, W.T.: Hypergraph regularity and the multidimensional Szemerédi theorem. *Ann. Math.* **166**(3), 897–946 (2007)
169. Grafakos, L.: Modern Fourier Analysis, vol. 250. Springer, Berlin (2009)
170. Grafakos, L., Li, X.: Uniform bounds for the bilinear Hilbert transforms I. *Ann. Math.* **159**(3), 889–933 (2004)
171. Granville, A.: Primes in intervals of bounded length. *Bull. Am. Math. Soc.* **52**(2), 171–222 (2015)
172. Granville, A., Soundararajan, K.: The spectrum of multiplicative functions. *Ann. Math.* **153**(2), 407–470 (2001)
173. Granville, A., Soundararajan, K.: An uncertainty principle for arithmetic sequences. *Ann. Math.* **165**(2), 593–635 (2007)
174. Green, B.: Roth’s theorem in the primes. *Ann. Math.* **161**(3), 1609–1636 (2005)
175. Green, B., Tao, T.: The primes contain arbitrarily long arithmetic progressions. *Ann. Math.* **167**(2), 481–547 (2008)
176. Green, B., Tao, T.: Linear equations in primes. *Ann. Math.* **171**(3), 1753–1850 (2010)
177. Green, B., Tao, T.: The Möbius function is strongly orthogonal to nilsequences. *Ann. Math.* **175**(2), 541–566 (2012)
178. Green, B., Tao, T., Ziegler, T.: An inverse theorem for the Gowers $U^{s+1}[N]$ -norm. *Ann. Math.* **176**(2), 1231–1372 (2012)
179. Grimmett, G.: Percolation, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 321. Springer, Berlin (1999)
180. Gruber, P.M.: Aspects of approximation of convex bodies. *Handbook of Convex Geometry*, Part A, pp. 319–345. Elsevier, Amsterdam (1993)
181. Guirao, A.J., Montesinos, V., Zizler, V., et al.: Open Problems in the Geometry and Analysis of Banach Spaces. Springer, Berlin (2016)
182. Hadamard, J.: Sur la distribution des zéros de la fonction $\zeta(s)$ et ses conséquences arithmétiques. *Bull. Soc. Math. Fr.* **24**(199–220), 2 (1896)
183. Häggström, O., Peres, Y., Steif, J.E.: Dynamical percolation. pp. 497–528. Association des Publications de l’Institut Henri Poincaré, Paris; Institut of Mathematical Statistics (IMS), Bethesda MD (1997)
184. Hales, T.: A proof of the Kepler conjecture. *Ann. Math.* **162**(3), 1065–1185 (2005)
185. Hales, T.: Historical overview of the Kepler conjecture. *The Kepler Conjecture*, pp. 65–82. Springer, Berlin (2011)
186. Hales, T.C., Adams, M., Bauer, G., Dang, T.D., Harrison, J., Le Truong, H., Kaliszyk, C., Magron, V., McLaughlin, S., Nguyen, T.T., et al.: A formal proof of the Kepler conjecture. *Forum Math. Pi* **5** (2017)
187. Hall, R.: Proof of a conjecture of Heath-Brown concerning quadratic residues. *Proc. Edinb. Math. Soc.* **39**(3), 581–588 (1996)
188. Hall, R.R.: A quantitative isoperimetric inequality in n -dimensional space. *J. reine angew. Math* **428**, 161–176 (1992)
189. Harper, A.J.: Sharp conditional bounds for moments of the Riemann zeta function. arXiv preprint [arXiv:1305.4618](https://arxiv.org/abs/1305.4618) (2013)

190. Hass, J., Hutchings, M., Schlaifly, R.: The double bubble conjecture. *Electron. Res. Announc. Am. Math. Soc.* **1**(3), 98–102 (1995)
191. Hass, J., Schlaifly, R.: Double bubbles minimize. *Ann. Math.* **151**(2), 459–515 (2000)
192. Håstad, J.: Some optimal inapproximability results. *J. ACM (JACM)* **48**(4), 798–859 (2001)
193. Heath-Brown, D.R.: The density of rational points on curves and surfaces. *Ann. Math.* **155**(2), 553–598 (2002)
194. Heinonen, J.: Lectures on Lipschitz analysis. University of Jyväskylä (2005)
195. Heinonen, J.: Nonsmooth calculus. *Bull. Am. Math. Soc.* **44**(2), 163–232 (2007)
196. Helfgott, H.A.: Growth and generation in $SL_2(\mathbb{Z}/p\mathbb{Z})$. *Ann. Math.* **167**(2), 601–623 (2008)
197. Helton, J.W.: “Positive” noncommutative polynomials are sums of squares. *Ann. Math.* **156**(2), 675–694 (2002)
198. Hertweck, M.: A counterexample to the isomorphism problem for integral group rings. *Ann. Math.* **154**(1), 115–138 (2001)
199. Higman, G.: The units of group-rings. *Proc. Lond. Math. Soc.* **2**(1), 231–248 (1940)
200. Higman, G., Neumann, B.H., Neuman, H.: Embedding theorems for groups. *J. Lond. Math. Soc.* **1**(4), 247–254 (1949)
201. Hilbert, D.: Mathematical problems. *Bull. Am. Math. Soc.* **8**(10), 437–479 (1902)
202. Hilbert, D.: Beweis für die Darstellbarkeit der ganzen Zahlen durch eine feste Anzahln ter Potenzen (Waring’sches Problem). *Mathematische Annalen* **67**(3), 281–300 (1909)
203. Hochman, M., Meyerovitch, T.: A characterization of the entropies of multidimensional shifts of finite type. *Ann. Math.* **171**(3), 2011–2038 (2010)
204. Hoory, S., Linial, N., Wigderson, A.: Expander graphs and their applications. *Bull. Am. Math. Soc.* **43**(4), 439–561 (2006)
205. Hopf, H.: Ein topologischer Beitrag zur reellen Algebra. *Commentarii Mathematici Helveticae* **13**(1), 219–239 (1941)
206. Host, B., Kra, B.: Nonconventional ergodic averages and nilmanifolds. *Ann. Math.* **161**(1), 397–488 (2005)
207. Hutchings, M., Morgan, F., Ritoré, M., Ros, A.: Proof of the double bubble conjecture. *Ann. Math.* **155**(2), 459–489 (2002)
208. Ireland, K., Rosen, M.: A Classical Introduction to Modern Number Theory, vol. 84. Springer Science & Business Media, Berlin (2013)
209. Izhboldin, O.T.: Fields of u -invariant 9. *Ann. Math.* **154**(3), 529–587 (2001)
210. Jaynes, E.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge (2003)
211. Johnson, W.B., Odell, E.: The diameter of the isomorphism class of a Banach space. *Ann. Math.* **162**(1), 423–437 (2005)
212. Jukna, S.: Extremal Combinatorics: With Applications in Computer Science. Springer Science & Business Media, Berlin (2011)
213. Kahn, J., Komlós, J., Szemerédi, E.: On the probability that a random ± 1 -matrix is singular. *J. Am. Math. Soc.* **8**(1), 223–240 (1995)
214. Kalai, G.: A Fourier-theoretic perspective on the Condorcet paradox and Arrow’s theorem. *Adv. Appl. Math.* **29**(3), 412–426 (2002)
215. Kaplansky, I.: Quadratic forms. *J. Math. Soc. Jpn.* **5**(2), 200–207 (1953)
216. Keith, S., Zhong, X.: The Poincaré inequality is an open ended condition. *Ann. Math.* **167**(2), 575–599 (2008)
217. Kempner, A.: Bemerkungen zum Waringschen Problem. *Mathematische Annalen* **72**(3), 387–399 (1912)
218. Kerckhoff, S., Masur, H., Smillie, J.: Ergodicity of billiard flows and quadratic differentials. *Ann. Math.* **124**(2), 293–311 (1986)
219. Kesten, H., Sidoravicius, V.: A shape theorem for the spread of an infection. *Ann. Math.* **167**(3), 701–766 (2008)
220. Kesten, H., Sidoravicius, V., et al.: The spread of a rumor or infection in a moving population. *Ann. Probab.* **33**(6), 2402–2462 (2005)

221. Khintchine, A.: Zwei Bemerkungen zu einer Arbeit des Herrn Perron. *Mathematische Zeitschrift* **22**(1), 274–284 (1925)
222. Khintchine, A.: Zur metrischen Theorie der diophantischen Approximationen. *Mathematische Zeitschrift* **24**(1), 706–714 (1926)
223. Khorunzhiy, O.: Estimates for moments of random matrices with Gaussian elements. In: Séminaire de probabilités XLI, pp. 51–92. Springer (2008)
224. Khot, S.: On the power of unique 2-prover 1-round games. In: Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing, pp. 767–775. ACM (2002)
225. Khot, S., Kindler, G., Mossel, E., O'Donnell, R.: Optimal inapproximability results for MAX-CUT and other 2-variable CSPs? *SIAM J. Comput.* **37**(1), 319–357 (2007)
226. Khot, S., Minzer, D., Safra, S.: Pseudorandom sets in grassmann graph have near-perfect expansion. *Electron. Colloq. Comput. Complex. (ECCC)* **25**, 6 (2018)
227. Klartag, B.: 5n Minkowski symmetrizations suffice to arrive at an approximate Euclidean ball. *Ann. Math.* **156**(3), 947–960 (2002)
228. Korkine, A., Zolotareff, G.: Sur les formes quadratiques. *Mathematische Annalen* **6**(3), 366–389 (1873)
229. Korkine, A., Zolotareff, G.: Sur les formes quadratiques positives. *Mathematische Annalen* **11**(2), 242–292 (1877)
230. Korte, B., Vygen, J.: Combinatorial Optimization, vol. 2. Springer, Berlin (2012)
231. Kozlovski, O., Shen, W., van Strien, S.: Density of hyperbolicity in dimension one. *Ann. Math.* **166**(1), 145–182 (2007)
232. Krammer, D.: Braid groups are linear. *Ann. Math.* **155**(1), 131–156 (2002)
233. Kuperberg, G.: Another proof of the alternative-sign matrix conjecture. *Int. Math. Res. Not.* **1996**(3), 139–150 (1996)
234. Kuperberg, G.: Symmetry classes of alternating-sign matrices under one roof. *Ann. Math.* **156**(3), 835–866 (2002)
235. Lam, T.Y.: The Algebraic Theory of Quadratic Forms. American Mathematical Society, Providence (1973)
236. Lam, T.Y.: Introduction to Quadratic Forms Over Fields, vol. 67. American Mathematical Society, Providence (2005)
237. Landman, B.M., Robertson, A.: Ramsey Theory on the Integers, vol. 73. American Mathematical Society, Providence (2014)
238. Last, Y.: Spectral theory of Sturm–Liouville operators on infinite intervals: a review of recent developments. *Sturm–Liouville Theory*, pp. 99–120. Springer, Berlin (2005)
239. Lehmer, D.H.: Factorization of certain cyclotomic functions. *Ann. Math.* **34**(3), 461–479 (1933)
240. Lemmermeyer, F.: Reciprocity Laws: From Euler to Eisenstein. Springer Science & Business Media, Berlin (2013)
241. Li, X., et al.: Uniform bounds for the bilinear Hilbert transforms. II. *Revista Matematica Iberoamericana* **22**(3), 1069–1126 (2006)
242. Liebeck, M.W., Shalev, A.: Diameters of finite simple groups: sharp bounds and applications. *Ann. Math.* **154**(2), 383–406 (2001)
243. Liggett, T.M.: Stochastic Interacting Systems: Contact, Voter and Exclusion Processes, vol. 324. Springer Science & Business Media, Berlin (2013)
244. Lindenstrauss, J., Preiss, D.: On Fréchet differentiability of Lipschitz maps between Banach spaces. *Ann. Math.* **157**(1), 257–288 (2003)
245. Lindenstrauss, J., Preiss, D., Tišer, J.: Fréchet Differentiability of Lipschitz Functions and Porous Sets in Banach Spaces (AM-179). Princeton University Press, Princeton (2012)
246. Linnik, J.V.: On the representation of large integers as sums of seven cubes. *Mat. Sb.* **12**, 218–224 (1943)
247. Littlewood, J.E.: Some Problems in Real and Complex Analysis. DC Heath (1968)
248. Ljunggren, W.: On the diophantine equation $x^2 + 4 = Ay^2$. *Norske Vid. Selsk. Forh., Trondheim* **24**, 82–84 (1951)

249. London, H., Finkelstein, R.: On Fibonacci and Lucas numbers which are perfect powers. *Fibonacci Quart* **7**(5), 476–481 (1969)
250. López, G.: Rational approximations, orthogonal polynomials and equilibrium distributions. *Orthogonal Polynomials and Their Applications*, pp. 125–157. Springer, Berlin (1988)
251. Lovász, L.: Normal hypergraphs and the perfect graph conjecture. *Discret. Math.* **2**(3), 253–267 (1972)
252. Lovasz, L., Grötschel, M., Schrijver, A.: *Geometric Algorithms and Combinatorial Optimization*. Springer, Berlin (1988)
253. Lubotzky, A.: Cayley graphs: eigenvalues, expanders and random walks. *Surveys in Combinatorics, 1995*. London Mathematical Society Lecture Note Series, pp. 155–190. Cambridge University Press, Cambridge (1995)
254. Luczak, T.: A note on the sharp concentration of the chromatic number of random graphs. *Combinatorica* **11**(3), 295–297 (1991)
255. Lyndon, R.C., Schupp, P.E.: *Combinatorial Group Theory*. Springer, Berlin (2015)
256. Lyubich, M.: Almost every real quadratic map is either regular or stochastic. *Ann. Math.* **156**(1), 1–78 (2002)
257. Maharam, D.: An algebraic characterization of measure algebras. *Ann. Math.* **48**(1), 154–167 (1947)
258. Mahler, K.: Zur Approximation algebraischer Zahlen. III. *Acta Math.* **62**(1), 91–166 (1933)
259. Maier, H., et al.: Primes in short intervals. *Mich. Math. J.* **32**(2), 221–225 (1985)
260. Maier, H., et al.: Small differences between prime numbers. *Mich. Math. J.* **35**(3), 323–344 (1988)
261. Marklof, J.: Pair correlation densities of inhomogeneous quadratic forms. *Ann. Math.* **158**(2), 419–471 (2003)
262. Martin, R., McMillen, W.: An elliptic curve over \mathbb{Q} with rank at least 24. *Number Theory Listserver* (2000)
263. Masur, H., et al.: Hausdorff dimension of the set of nonergodic foliations of a quadratic differential. *Duke Math. J.* **66**(3), 387–442 (1992)
264. Matoušek, J.: *Lectures on Discrete Geometry*, vol. 212. Springer, New York (2002)
265. Mauduit, C., Rivat, J.: Sur un problème de Gelfond: la somme des chiffres des nombres premiers. *Ann. Math.* **171**(3), 1591–1646 (2010)
266. Mazur, B., Stein, W.: *Prime Numbers and the Riemann Hypothesis*. Cambridge University Press, Cambridge (2016)
267. Meeks III, W., Pérez, J.: The classical theory of minimal surfaces. *Bull. Am. Math. Soc.* **48**(3), 325–407 (2011)
268. Meeks III, W.H., Rosenberg, H.: The uniqueness of the helicoid. *Ann. Math.* **161**(2), 727–758 (2005)
269. Melas, A.D.: The best constant for the centered Hardy-Littlewood maximal inequality. *Ann. Math.* **157**(2), 647–688 (2003)
270. Mendel, M., Naor, A.: Ramsey partitions and proximity data structures. *J. Eur. Math. Soc.* **9**(2), 253–275 (2007)
271. Mendel, M., Naor, A.: Metric cotype. *Ann. Math.* **168**(1), 247–298 (2008)
272. Mendel, M., Naor, A.: Ultrametric skeletons. *Proc. Natl. Acad. Sci.* **110**(48), 19256–19262 (2013)
273. Merkur'ev, A.: Kaplansky conjecture in the theory of quadratic forms. *J. Sov. Math.* **57**(6), 3489–3497 (1991)
274. Merkur'ev, A.S.: Simple algebras and quadratic forms. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya* **55**(1), 218–224 (1991)
275. Meusnier, J.B.: Mémoire sur la courbure des surfaces. *Mem des savants étrangers* **10**(1776), 477–510 (1785)
276. Miller, G.L.: Riemann's hypothesis and tests for primality. *J. Comput. Syst. Sci.* **13**(3), 300–317 (1976)
277. Milnor, J.W.: *Dynamics in One Complex Variable*, vol. 160. Springer, Berlin (2006)
278. Moore, C., Mertens, S.: *The Nature of Computation*. OUP, Oxford (2011)

279. Moreira, C.G.: Geometric properties of the Markov and Lagrange spectra. arXiv preprint [arXiv:1612.05782](https://arxiv.org/abs/1612.05782) (2016)
280. Moreira, C.G.: Geometric properties of the Markov and Lagrange spectra. Ann. Math. **188**(1), 145–170 (2018)
281. Moreira, C.G., Yoccoz, J.C.: Stable intersections of regular Cantor sets with large Hausdorff dimensions. Ann. Math. **154**(1), 45–96 (2001)
282. Morgan, F.: Geometric Measure Theory: A Beginner’s Guide. Academic Press, Cambridge (2016)
283. Mossel, E., ODonnell, R., Oleszkiewicz, K.: Noise stability of functions with low influences: invariance and optimality. Ann. Math. **171**(1), 295–341 (2010)
284. Motzkin, T.S.: The arithmetic-geometric inequality. In: Inequalities (Proceedings of a symposium held at Wright-Patterson Air Force Base, Ohio, 1965) pp. 205–224 (1967)
285. Mukhin, E., Tarasov, V., Varchenko, A.: The B. and M. Shapiro conjecture in real algebraic geometry and the Bethe ansatz. Ann. Math. **170**(2), 863–881 (2009)
286. Mullen, G.L., Panario, D.: Handbook of Finite Fields. Chapman and Hall/CRC, Boca Raton (2013)
287. Musin, O.R.: The kissing number in four dimensions. Ann. Math. **168**(1), 1–32 (2008)
288. Myasnikov, G.B.A.G., Shpilrain, V.: Open problems in combinatorial group theory. In: Combinatorial and Geometric Group Theory: AMS Special Session, Combinatorial Group Theory, 4–5 November 2000, New York (AMS Special Session, Computational Group Theory, 28–29 April 2001, Hoboken, New Jersey), vol. 296, p. 1 (2002)
289. Niculescu, C., Persson, L.E.: Convex Functions and Their Applications. Springer, Berlin (2006)
290. Novikov, P.S.: Algorithmic Unsolvability of the Word Problem in Group Theory. Trydu Mat. Inst. Stelkov **44**, (1958)
291. Odlyzko, A.M., Sloane, N.J.: New bounds on the number of unit spheres that can touch a unit sphere in n dimensions. J. Comb. Theory Ser. A **26**(2), 210–214 (1979)
292. Okada, S.: Enumeration of symmetry classes of alternating sign matrices and characters of classical groups. J. Algebr. Comb. **23**(1), 43–69 (2006)
293. Ono, K.: Distribution of the partition function modulo m . Ann. Math. **151**(1), 293–307 (2000)
294. Ono, K., et al.: Unearthing the visions of a master: harmonic Maass forms and number theory. Curr. Dev. Math. **2008**, 347–454 (2009)
295. Osin, D.: Small cancellations over relatively hyperbolic groups and embedding theorems. Ann. Math. **172**(1), 1–39 (2010)
296. Palis, J.: Homoclinic orbits, hyperbolic dynamics and dimension of Cantor sets. Contemp. Math. **58**(26), 203–216 (1987)
297. Palis, J.: A global view of dynamics and a conjecture on the denseness of finitude of attractors. Astérisque **261**(xiixiv), 335–347 (2000)
298. Palis, J.: A global perspective for non-conservative dynamics. Ann. Inst. Henri Poincaré, Anal. Non Linéaire **22**(4), 485–507 (2005)
299. Palis, J., Takens, F.: Hyperbolicity and Sensitive Chaotic Dynamics at Homoclinic Bifurcations: Fractal Dimensions and Infinitely Many Attractors in Dynamics. Cambridge Studies in Advanced Mathematics, vol. 35. Cambridge University Press, Cambridge (1995)
300. Permantle, R.: Choosing a spanning tree for the integer lattice uniformly. Ann. Probab. **19**(4), 1559–1574 (1991)
301. Petersen, C.L.: Local connectivity of some Julia sets containing a circle with an irrational rotation. Acta Math. **177**(2), 163–224 (1996)
302. Petersen, C.L., Zakeri, S.: On the Julia set of a typical quadratic polynomial with a Siegel disk. Ann. Math. **159**(1), 1–52 (2004)
303. Pethő, A.: Diophantine properties of linear recursive sequences II. Acta Math. Acad. Paed. Nyiregyháziensis **17**, 81–96 (2001)
304. Pfender, F., Ziegler, G.M.: Kissing numbers, sphere packings, and some unexpected proofs. Not.-Am. Math. Soc. **51**, 873–883 (2004)

305. Pietsch, A.: Theorie der Operatorenideale: Zusammenfassung. Friedrich-Schiller Universität (1972)
306. Pila, J.: Density of integral and rational points on varieties. *Astérisque* **228**, 183–187 (1995)
307. Pólya, G., Pólya, G., Szegő, G.: Isoperimetric Inequalities in Mathematical Physics. Princeton University Press, Princeton (1951)
308. Pólya, G., Schur, J.: Über zwei Arten von Faktorenfolgen in der Theorie der algebraischen Gleichungen. *Journal für die reine und angewandte Mathematik* **144**, 89–113 (1914)
309. Polymath, D.H.J.: Variants of the Selberg sieve, and bounded intervals containing many primes. *Res. Math. Sci.* **1**(1), 12 (2014)
310. Prasolov, V.V.: Polynomials, vol. 11. Springer Science & Business Media, Berlin (2009)
311. Rabin, M.O.: Probabilistic algorithm for testing primality. *J. Number Theory* **12**(1), 128–138 (1980)
312. Rakhmanov, E.A.: Bounds for polynomials with a unit discrete norm. *Ann. Math.* **165**(1), 55–88 (2007)
313. Ramachandra, K.: Some remarks on the mean value of the Riemann zeta-function and other Dirichlet series 1. *Hardy-Ramanujan J.* **1**, 1–15 (1978)
314. Ramsey, F.P.: On a problem of formal logic. *Proc. Lond. Math. Soc.* **2**(1), 264–286 (1930)
315. Razumov, A.V., Stroganov, Y.G.: Enumeration of quarter-turn-symmetric alternating-sign matrices of odd order. *Theor. Math. Phys.* **149**(3), 1639–1650 (2006)
316. Razumov, A.V., Stroganov, Y.G.: Enumerations of half-turn-symmetric alternating-sign matrices of odd order. *Theor. Math. Phys.* **148**(3), 1174–1198 (2006)
317. Reichardt, B.W., Heilmann, C., Lai, Y.Y., Spielman, A.: Proof of the double bubble conjecture in R^4 and certain higher dimensional cases. *Pac. J. Math.* **208**(2), 347–366 (2003)
318. Reingold, O.: Undirected connectivity in log-space. *J. ACM (JACM)* **55**(4), 17 (2008)
319. Reingold, O., Vadhan, S., Wigderson, A.: Entropy waves, the zig-zag graph product, and new constant-degree expanders. *Ann. Math.* **155**(1), 157–187 (2002)
320. Ricci, G.: Sull'andamento della differenza di numeri primi consecutivi (1954)
321. Richert, H.E.: Über Zerfällungen in ungleiche Primzahlen. *Mathematische Zeitschrift* **52**(1), 342–343 (1949)
322. Riemann, B.: Ueber die Anzahl der Primzahlen unter einer gegebenen Grosse. *Ges. Math. Werke und Wissenschaftlicher Nachlaß* **2**, 145–155 (1859)
323. Robbins, D.P.: The Story of 1, 2, 7, 42, 429, 7436. *Math. Intell.* **13**(2), 12–19 (1991)
324. Roche-Newton, O., Rudnev, M., Shkredov, I.D.: New sum-product type estimates over finite fields. *Adv. Math.* **293**, 589–605 (2016)
325. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (2015)
326. Roth, K.: On certain sets of integers. *J. Lond. Math. Soc.* **1**(1), 104–109 (1953)
327. Roth, K.: Remark concerning integer sequences. *Acta Arith.* **3**(9), 257–260 (1964)
328. Roy, D.: Approximation to real numbers by cubic algebraic integers. II. *Ann. Math.* **158**(3), 1081–1087 (2003)
329. Rudelson, M.: Invertibility of random matrices: norm of the inverse. *Ann. Math.* **168**(2), 575–600 (2008)
330. Rudelson, M., Vershynin, R.: Combinatorics of random processes and sections of convex bodies. *Ann. Math.* **164**(2), 603–648 (2006)
331. Sanders, T.: On Roth's theorem on progressions. *Ann. Math.* **174**(1), 619–636 (2011)
332. Sapir, M.V., Birget, J.C., Rips, E.: Isoperimetric and isodiametric functions of groups. *Ann. Math.* **156**(2), 345–466 (2002)
333. Savin, O.: Regularity of flat level sets in phase transitions. *Ann. Math.* **169**(1), 41–78 (2009)
334. Schinzel, A., Zassenhaus, H., et al.: A refinement of two theorems of Kronecker. *Michigan Math. J.* **12**, 81–85 (1965)
335. Schmidt, W.M.: Number fields of given degree and bounded discriminant. *Astérisque* **228**(4), 189–195 (1995)
336. Schmidt, W.M.: Diophantine Approximation. Lecture Notes in Mathematics, vol. 785. Springer Science & Business Media, Berlin (1996)

337. Schramm, O.: Conformally invariant scaling limits: an overview and a collection of problems. Selected Works of Oded Schramm, pp. 1161–1191. Springer, Berlin (2011)
338. Schramm, O., Steif, J.E.: Quantitative noise sensitivity and exceptional times for percolation. *Ann. Math.* **171**(2), 619–672 (2010)
339. Schütte, K., van der Waerden, B.L.: Das problem der dreizehn Kugeln. *Mathematische Annalen* **125**(1), 325–334 (1952)
340. Schwartz, R.E.: Obtuse triangular billiards II: One hundred degrees worth of periodic trajectories. *Exp. Math.* **18**(2), 137–171 (2009)
341. Schwarz, H.A.: Beweis des Satzes, dass die Kugel kleinere Oberfläche besitzt, als jeder andere Körper gleichen Volumens. Nachrichten von der Königl. Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen **1884**, 1–13 (1884)
342. Sebbar, A., Falliero, T.: Equilibrium point of Greens function for the annulus and Eisenstein series. *Proc. Am. Math. Soc.* **135**(2), 313–328 (2007)
343. Segal, D.: Words: Notes on Verbal Width in Groups, vol. 361. Cambridge University Press, Cambridge (2009)
344. Semmes, S.: Some Novel Types of Fractal Geometry. Oxford University Press, Oxford (2001)
345. Shalev, A.: Word maps, conjugacy classes, and a noncommutative Waring-type theorem. *Ann. Math.* **170**(3), 1383–1416 (2009)
346. Silverman, J.H.: The Arithmetic of Elliptic Curves, vol. 106. Springer Science & Business Media, Berlin (2009)
347. Skolem, T.: Diophantische gleichungen. J. Springer (1938)
348. Smale, S.: Mathematical problems for the next century. *Math. Intell.* **20**(2), 7–15 (1998)
349. Smyth, C.: The Mahler measure of algebraic numbers: a survey. Number Theory and Polynomials. London Mathematical Society Lecture Note Series, p. 322. Cambridge University Press, Cambridge (2008)
350. Sodin, S.: The spectral edge of some random band matrices. *Ann. Math.* **172**(3), 2223–2251 (2010)
351. Solynin, A.: A note on equilibrium points of Greens function. *Proc. Am. Math. Soc.* **136**(3), 1019–1021 (2008)
352. Solynin, A.Y., Zalgaller, V.A.: An isoperimetric inequality for logarithmic capacity of polygons. *Ann. Math.* **159**(1), 277–303 (2004)
353. Sottile, F.: Real Schubert calculus: polynomial systems and a conjecture of Shapiro and Shapiro. *Exp. Math.* **9**(2), 161–182 (2000)
354. Soundararajan, K.: Moments of the Riemann zeta function. *Ann. Math.* **170**(2), 981–993 (2009)
355. Spitzer, F.: Electrostatic capacity, heat flow, and Brownian motion. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **3**(2), 110–121 (1964)
356. Spitzer, F.: Interaction of Markov processes. *Adv. Math.* **5**(2), 246–290 (1970)
357. Spitzer, F.: Principles of Random Walk. Graduate Texts in Mathematics, vol. 34. Springer Science & Business Media, Berlin (2013)
358. Spohn, H.: Large Scale Dynamics of Interacting Particles. Springer Science & Business Media, Berlin (2012)
359. Sprague, R.: Über zerlegungen in ungleiche Quadratzahlen. *Mathematische Zeitschrift* **51**(3), 289–290 (1948)
360. Stein, E.M.: Singular Integrals and Differentiability Properties of Functions (PMS-30), vol. 30. Princeton University Press, Princeton (2016)
361. Stevenhagen, P.: The number of real quadratic fields having units of negative norm. *Exp. Math.* **2**(2), 121–136 (1993)
362. Stewart, I.: Mathematics: Where drunkards hang out. *Nature* **413**(6857), 686 (2001)
363. Szabados, T.: An elementary introduction to the Wiener process and stochastic integrals. arXiv preprint [arXiv:1008.1510](https://arxiv.org/abs/1008.1510) (2010)
364. Szarek, S.J.: Condition numbers of random matrices. *J. Complex.* **7**(2), 131–149 (1991)
365. Szemerédi, E.: On sets of integers containing no k elements in arithmetic progression. *Acta Arith.* **27**, 299–345 (1975)

366. Szemerédi, E., Vu, V.H.: Finite and infinite arithmetic progressions in sumsets. *Ann. Math.* **163**(1), 1–35 (2006)
367. Talagrand, M.: Type, infratype and the Elton-Pajor theorem. *Inventiones mathematicae* **107**(1), 41–59 (1992)
368. Talagrand, M.: Maharam’s problem. *Ann. Math.* **168**(3), 981–1009 (2008)
369. Tao, T.: An Introduction to Measure Theory. American Mathematical Society, Providence, RI (2011)
370. Tao, T.: Topics in Random Matrix Theory, vol. 132. American Mathematical Society, Providence (2012)
371. Tao, T., Croot III, E., Helfgott, H.: Deterministic methods to find primes. *Math. Comput.* **81**(278), 1233–1246 (2012)
372. Tao, T., Vu, V.H.: Inverse Littlewood-Offord theorems and the condition number of random discrete matrices. *Ann. Math.* **169**(2), 595–632 (2009)
373. Tao, T., Ziegler, T.: The primes contain arbitrarily long polynomial progressions. *Acta Math.* **201**(2), 213–305 (2008)
374. Thiele, C.: A uniform estimate. *Ann. Math.* **156**(2), 519–563 (2002)
375. Thomason, A.: Pseudo-random graphs. North-Holland Mathematics Studies, vol. 144, pp. 307–331. Elsevier, Amsterdam (1987)
376. Thunder, J.L.: Decomposable form inequalities. *Ann. Math.* **153**(3), 767–804 (2001)
377. Tikhomirov, K.: Singularity of random bernoulli matrices. arXiv preprint [arXiv:1812.09016](https://arxiv.org/abs/1812.09016) (2018)
378. Titchmarsh, E.C.T., Titchmarsh, E.C., Heath-Brown, D.R., et al.: The Theory of the Riemann Zeta-Function. Oxford University Press, Oxford (1986)
379. Todorcevic, S.: Introduction to Ramsey Spaces (am-174). Princeton University Press, Princeton (2010)
380. Tomczak-Jaegermann, N.: Banach–Mazur Distances and Finite-Dimensional Operator Ideals, vol. 38. Longman Sc & Tech (1989)
381. Tracy, C.A., Widom, H.: Integral formulas for the asymmetric simple exclusion process. *Commun. Math. Phys.* **279**(3), 815–844 (2008)
382. Ulmer, D.: Elliptic curves with large rank over function fields. *Ann. Math.* **155**(1), 295–315 (2002)
383. van Beijeren, H., Kutner, R., Spohn, H.: Excess noise for driven diffusive systems. *Phys. Rev. Lett.* **54**(18), 2026 (1985)
384. van den Berg, M., Bolthausen, E., den Hollander, F.: Moderate deviations for the volume of the Wiener sausage. *Ann. Math.* **153**(2), 355–406 (2001)
385. van den Berg, M., Bolthausen, E., den Hollander, F.: On the volume of the intersection of two Wiener sausages. *Ann. Math.* **159**(2), 741–782 (2004)
386. van der Corput, J.G.: Über Summen von Primzahlen und Primzahlquadraten. *Mathematische Annalen* **116**(1), 1–50 (1939)
387. Vaserstein, L.: Polynomial parametrization for the solutions of Diophantine equations and arithmetic groups. *Ann. Math.* **171**(2), 979–1009 (2010)
388. Vaughan, R.C.: On pairs of additive cubic equations. *Proc. Lond. Math. Soc.* **3**(2), 354–364 (1977)
389. Veech, W.A.: The metric theory of interval exchange transformations I. Generic spectral properties. *Am. J. Math.* **106**(6), 1331–1359 (1984)
390. Venkatesh, A.: The work of Einsiedler, Katok and Lindenstrauss on the Littlewood conjecture. *Bull. Am. Math. Soc.* **45**(1), 117 (2008)
391. Viana, M.: Ergodic theory of interval exchange maps. *Revista Matemática Complutense* **19**(1), 7–100 (2006)
392. Viazovska, M.S.: The sphere packing problem in dimension 8. *Ann. Math.* **185**(3), 991–1015 (2017)
393. Vishik, A.: Fields of u -invariant $2^r + 1$. Algebra, Arithmetic, and Geometry, pp. 661–685. Springer, Berlin (2009)

394. Vitali, G.: Sul problema della misura dei Gruppi di punti di una retta: Nota. Tip. Gamberini e Parmeggiani (1905)
395. Vorobiev, N.N.: Fibonacci Numbers. Birkhäuser, Basel (2012)
396. Wagner, D.: Multivariate stable polynomials: theory and applications. Bull. Am. Math. Soc. **48**(1), 53–84 (2011)
397. Walters, P.: An Introduction to Ergodic Theory, vol. 79. Springer Science & Business Media, Berlin (2000)
398. Weber, M., Hoffman, D., Wolf, M.: An embedded genus-one helicoid. Ann. Math. **169**(2), 347–448 (2009)
399. Weierstrass, K.: Mathematische werke: Abhandlungen 2, vol. 2. Georg Olms (1895)
400. Weil, A.: Number Theory: An Approach Through History. From Hammurapi to Legendre. Springer Science & Business Media, Berlin (2006)
401. West, D.B., et al.: Introduction to Graph Theory, vol. 2. Prentice hall, Upper Saddle River (2001)
402. Whitney, H.: Analytic extensions of differentiable functions defined in closed sets. Trans. Am. Math. Soc. **36**(1), 63–89 (1934)
403. Wieferich, A.: Beweis des Satzes, daß sich eine jede ganze Zahl als Summe von höchstens neun positiven Kuben darstellen läßt. Mathematische Annalen **66**(1), 95–101 (1908)
404. Williamson, D.P., Shmoys, D.B.: The Design of Approximation Algorithms. Cambridge University Press, Cambridge (2011)
405. Yau, H.T.: $(\ln t)^{2/3}$ law of the Two Dimensional Asymmetric Simple Exclusion Process. Ann. Math. **159**(1), 377–405 (2004)
406. Zannier, U.: Some Problems of Unlikely Intersections in Arithmetic and Geometry. Annals of Mathematics Studies, vol. 181. Princeton University Press, Princeton (2012)
407. Zeilberger, D.: Proof of the alternating sign matrix conjecture. Electron. J. Combin. **3**(2), R13 (1996)
408. Zhang, Y.: Bounded gaps between primes. Ann. Math. **179**(3), 1121–1174 (2014)

Author Index

A

- Abel, N.H., 23
Achilioptas, D., 190
Adamczewski, B., 223, 224
Agrawal, M., 148
Alon, N., 327, 328
Archimedes, 48
Arrow, K.J., 359
Artin, E., 35
Artstein, S., 132, 133
Artstein-Avidan, S., 321
Atkin, A.O.L., 361, 363
Avila, A., 163, 231, 325
Azbel, M.Ya., 325

B

- Bachet (C.G. Bachet de Méziriac), 192
Baker, R.C., 249
Balázs, M., 374
Banach, S., 72, 82, 171, 172, 298, 299
Bartal, Y., 179
Behrend, F.A., 166
Behrend, R.E., 76
Belius, D., 141
Benjamini, I., 143, 144
Beresnevich, V., 214, 215, 246
Berge, C., 204
Bertrand, J., 192
Bhargava, M., 136, 137, 183, 403
Bigelow, S., 37
Birch, B., 45
Birget, J.-C., 62, 65, 66
Birkhoff, G.D., 376–378
Blichfeldt, H.F., 349
Bloom, T.F., 168
Bolthausen, E., 5, 126

Bombieri, E., 334

- Boole, G., 300
Borcea, J., 332
Borel, É., 205, 243, 297
Borel, F.E.J.E., 107
Borwein, P., 242, 243, 278, 279
Bourgain, J., 166, 168, 272
Brändén, P., 332
Bringmann, K., 363
Brown, R., 125, 140
Brúdern, J., 250
Buczolich, Z., 378
Bugeaud, Y., 201, 223, 224
Burau, W., 37

C

- Calabi, E., 258
Calderón, A., 129
Cantor, G.F.L.P., 15–18
Catalan, E.C., 160
Cauchy, A.-L., 132
Cayley, A., 271
Chang, M.-C., 98
Chebyshev, P.L., 118, 192
Cheng, Zh., 100
Cheung, Y., 105, 106
Chudnovsky, M., 204
Cohn, H., 94, 95, 349
Colding, T., 258
Collins, D.J., 65
Conlon, D., 342
Cook, S., 174
Coppersmith, D., 220
Coxeter, H.S.M., 282
Croot III, E.S., 87, 88

D

- Davenport, H., 108, 250, 403
 De Giorgi, E., 306
 Del Pino, M., 306
 Delsarte, P., 282
 Dembo, A., 140, 141
 Den Hollander, F., 5, 126
 Dickinson, D., 246
 Dinur, I., 175, 176
 Diophantus, 231, 232, 350
 Dirichlet, J.P.G.L., 205, 212, 243, 246, 333
 Dobrowolski, E., 242, 243
 Duffin, R.J., 214
 Dugger, D., 234, 235
 Dyson, F.J., 362, 363

E

- Einsiedler, M., 207
 Elkies, N., 94, 95
 Ellenberg, J., 197, 198
 Elliott, P.D.T.A., 334
 Ennepet, A., 159, 258
 Eratosthenes, 144
 Erdélyi, T., 279
 Erdős, P., 86, 88, 97, 165, 166, 194, 288–290,
 332, 341, 342
 Eremenko, A., 33, 34
 Euclid, 224
 Euler, L., 114, 158, 159, 232, 313, 344, 404
 Evans, R.J., 341

F

- Fatou, P., 115
 Fefferman, C.L., 157, 308–310
 Ferenczi, S., 222
 Ferguson, R., 279
 Fermat, P., 146, 404
 Ferrari, P.A., 374
 Fibonacci (Leonardo of Pisa), 198
 Figalli, A., 297
 Finkelstein, R., 199
 Fischer, I., 76
 Fontes, L.R.G., 374
 Ford, K., 289, 290
 Forni, G., 231
 Fourier, J.-B.J., 94, 175
 Fouvry, É., 408
 Fraenkel, L.E., 296
 Frind, M., 260
 Fusco, N., 297

G

- Gabrielov, A., 33, 34
 Gamburd, A., 272
 Gauss, C.F., 312, 349
 Gelfond, A., 381, 382
 Germain, S., 386–388
 Goldberg, L.R., 33, 34
 Goldston, D.A., 332, 333, 335
 Gowers, W.T., 72, 73, 253, 254
 Grafakos, L., 129, 130
 Graham, R.L., 86, 88
 Granville, A., 6, 9, 10, 226–228
 Green, B., 167, 168, 261, 262, 387–389
 Gregory, D., 281

H

- Hadamard, J.S., 58, 344
 Halberstam, H., 334
 Hales, T., 186, 187
 Hall, R.R., 6, 296
 Hardy, G.H., 91, 92, 332
 Harper, A., 346
 Hasse, H., 249
 Håstad, J., 175
 Hausdorff, F., 17, 104, 207, 214
 Heath-Brown, D.R., 53
 Heilbronn, H., 403
 Helfgott, H., 269
 Helton, J.W., 69
 Hertweck, M., 22
 Higman, G., 21, 22
 Hilbert, D., 129, 344, 350
 Hoene-Wroński, J.M., 338
 Hoffman, D., 314
 Hohman, M., 393
 Hopf, H., 233, 234
 Host, B., 152–154
 Hutchings, M., 49

I

- Isaksen, D., 234, 235
 Izboldin, O., 30

J

- Jitomirskaya, S., 325
 Johnson, W., 171
 Jordan, C., 115
 Julia, G., 112

K

- Kac, M., 325

Kahn, J., 317
 Kalai, G., 359
 Kaplansky, I., 29
 Katok, A., 207
 Kayal, N., 148
 Keith, S., 265
 Kempner, A., 350
 Kepler, J., 186
 Kerckhoff, S., 102
 Kesten, H., 143, 144, 275
 Khinchin, A.Ya., 213, 244, 246
 Khorunzhiy, O., 411
 Khot, S., 359
 Kindler, G., 359
 Kistler, N., 141
 Klartag, B., 78, 80, 309, 310
 Klüners, J., 408
 Knot, S., 176
 Koltchinskii, V., 210
 Konvalinka, M., 76
 Korkine, A., 349
 Kowalczyk, M., 306
 Kozlovski, O., 239
 Kra, B., 152–154
 Krammer, D., 37, 38
 Kumar, A., 349
 Kuperberg, G., 74–77

L

Lagrange, J.-L., 192, 232, 233, 350
 Lebesgue, H.L., 84, 90, 91, 153, 200, 213
 Lebowitz, J.L., 100
 Leech, J., 349
 Legendre, A.-M., 320, 321
 Lehmer, D.H., 242
 Levin, L., 174
 Lewis, D.J., 250
 Liebeck, M., 26
 Lindenstrauss, E., 207
 Lindenstrauss, J., 84
 Linial, N., 179
 Linnik, J.V., 350
 Liouville, J., 101, 244
 Lipschitz, R.O.S., 83, 264
 Littlewood, J.E., 91, 92, 206, 278, 332
 Li, X., 129, 130
 Ljunggren, W., 199
 Lockhart, R., 279
 London, H., 199
 Lucas, F.E.A., 199
 Lyubich, M., 58

M

Maggi, F., 297
 Maharam, D., 301
 Mahler, K., 13, 242
 Maier, H., 226, 332
 Manduit, C., 222
 Marklof, J., 101, 102
 Masur, H., 102, 104, 105
 Mathieu, É.L., 324
 Mauduit, C., 382
 Mauldin, D., 378
 Maynard, J., 335
 Mazur, B.C., 171
 Meeks III, W.H., 160
 Melas, A.D., 92
 Mendel, M., 179, 286
 Merkurev, A., 29
 Meyerovitch, T., 393
 Mignotte, M., 201
 Miller, G.L., 147, 148
 Milman, V.D., 132, 133, 321
 Minicozzi, W., 258
 Minkowski, H., 77, 79, 249
 Moreira, C.J., 17, 18, 163
 Morgan, F., 49
 Mossel, E., 359, 360
 Mossinghoff, M., 242, 243
 Motzkin, T., 66
 Mukhin, E., 338
 Musin, O., 282

N

Nagell, T., 200
 Naor, A., 179, 190, 286
 Newton, I., 53, 281, 375
 Ngô, B.C., v

O

O'Brien, J.N., 361
 Odell, E., 171
 O'Donnell, R., 359, 360
 Okada, S., 76
 Oleszkiewicz, K., 360
 Ol'shanskii, A.Yu., 65, 66
 Ono, K., 362, 363
 Osin, D., 397

P

Painlevé, P., 58
 Paley, R., 341
 Palis, J., 16

- Pell, J., 404
 Pemantle, R., 143
 Perelman, G.Y., vi
 Peres, Y., 140, 141, 143, 144
 Petersen, C.L., 112, 114, 115
 Pethő, A., 199
 Pietsch, A., 133
 Pila, J., 52
 Pintz, J., 332—335
 Poincaré, J.H., vi
 Poisson, S.D., 98, 273
 Pollard, D., 210
 Pólya, G., 119, 124, 330
 Pratelli, A., 297
 Preiss, D., 84
 Pulham, J.R., 341
 Pythagoras, 21, 106, 211, 367, 389
- R**
 Rabin, M.O., 147, 148
 Rakhmanov, E.A., 220, 221
 Ramachandra, K., 345, 346
 Ramanujan, S., 361, 362, 389
 Ramsey, F.P., 73, 341
 Razumov, A.V., 76
 Reingold, O., 39, 41, 42
 Ricci, G., 332
 Richert, H.-E., 192
 Riemann, G.F.B., 332, 342, 343
 Rips, E., 62, 65, 66
 Rivat, J., 382
 Rivlin, T.J., 220
 Robbins, D., 74, 76
 Robertson, N., 204
 Ros, A., 49
 Rosenberg, H., 160
 Rosen, J., 140, 141
 Roth, K.F., 166, 167, 251
 Ritiré, M., 49
 Roy, D., 109
 Rudelson, M., 211, 294, 317, 318
 Rumsey, H., 74
- S**
 Safra, S., 175, 176
 Sanders, T., 168
 Sapir, M.V., 62, 65, 66
 Savin, O., 306
 Saxena, N., 148
 Schaeffer, A.C., 214
 Schinzel, A., 242
- Schmidt, W., 108
 Schramm, O., 143, 144, 366
 Schur, I., 330
 Schütte, K., 281
 Schwartz, L., 128
 Schwartz, R., 106
 Schwarz, H.A., 48, 132
 Seppäläinen, T., 374
 Seymour, P., 204
 Shalev, A., 26, 353
 Shapira, A., 327, 328
 Shapiro, B., 34, 338
 Shapiro, M., 34, 338
 Sheehan, J., 341
 Shen, W., 239
 Sidoravicius, V., 275
 Siegel, C.L., 115
 Siksek, S., 201
 Sinai, Ya., 163
 Skolem, T., 369
 Smale, S., 238
 Smillie, J., 102
 Sodin, S., 411, 412
 Solynin, A., 119
 Soundararajan, K., 6, 9, 10, 226–228, 346
 Spitzer, F., 4
 Sprague, R., 192
 Steif, J., 366
 Stevenhagen, P., 408
 Stroganov, Yu.G., 76
 Sudakov, B., 327, 328
 Swinnerton-Dyer, P., 45, 363
 Szarek, S.J., 132, 133
 Szegő, G., 119
 Szekeres, G., 341
 Szemerédi, E., 97, 151, 193, 194, 252, 253
- T**
 Talagrand, M., 301
 Tao, T., 168, 261, 262, 317, 318, 387–389
 Tarasov, V., 338
 Tarski, A., 298, 299
 Thomason, A., 341, 342
 Thomas, R., 204
 Thue, A., 93
 Thunder, J.L., 12–14
 Turán, P., 165, 166
- U**
 Ulmer, D., 45, 46

V

- Vadhan, S., 39, 41, 42
Van den Berg, M., 5, 126
Van der Corput, J.G., 167, 260
Van der Waerden, B.L., 281
Van Strien, S., 239
Varchenko, A., 338
Vaserstein, L., 369, 370
Vaughan, R.C., 249, 250
Velani, S., 214, 215, 246
Venkatesh, A., 197, 198
Vershynin, R., 211
Viazovska, M., 95
Vinogradov, A.I., 334
Vitali, G., 298, 299
Vu, V., 193, 194, 317, 318

W

- Waring, E., 350
Weber, M., 314
Weierstrass, K.T.W., 83
Wei, J., 306

Whitney, H., 157

- Widgerson, A., 39, 41, 42
Wieferich, A., 350
Wiener, N., 4, 125
Wolf, M., 314

Wooley, T., 250

Y

- Yau, H.-T., 122, 123
Yıldırım, C., 332—335
Yoccoz, J.-C., 17, 18

Z

- Zakeri, S., 112, 114, 115
Zalgaller, Z., 119
Zassenhaus, H., 242
Zeilberger, D., 74
Zeitouni, O., 140, 141
Zhang, Y., 335
Zhong, X., 265
Zolotareff, G., 349

Index

A

Abelian group, 23, 133

Absolutely continuous submeasure, 301

Absolutely integrable function, 378

Absolute value, 241, 343

Additive inverse, 194

Admissible

colouring, 392

function, 95

sequence, 387

Affine-linear

form, 384

transformation, 383

Algebraic

integer, 108, 181, 401

number, 222, 399

number field, 182

Algorithm, 61

polynomial, 61

Almost all, 206, 246, 376

Almost every, 58, 213, 231

f -almost every, 214

Almost Mathieu operator, 324, 325

Almost surely, 188, 275

Alternating-Sign Matrix (ASM), 73

Annals of Mathematics, v

Approximation theory, 217

3-term Arithmetic Progression (3AP), 164

Arithmetic progression, 96, 251, 260, 333,

384

generalized, 151

of length 3, 260

of length k , 260

Arithmetic sequence, 227

Arrow's impossibility theorem, 359

Associativity, 180, 194

Asymmetric Simple Exclusion Process (ASEP), 372

Attracting

cycle, 56, 161

fixed point, 54, 236

periodic point, 237

Average value, 3

Axiom of choice, 298

B

Ball, 79, 112

Banach space, 72, 82, 172

Banach–Tarski paradox, 298, 299

B. and M. Shapiro conjecture, 34, 338

Base of number system, 223

Basis, 335, 337

of a number field, 400

Beresnevich–Velani conjecture, 214

Berge's conjecture, 204

Bertrand's postulate, 192

Bijection, 18

Bilinear Hilbert transform, 129

Binary cubic form, 136

Binomial coefficients, 233

Birch and Swinnerton-Dyer conjecture, 45

Birkhoff's ergodic theorem, 376–378

Bombieri–Vinogradov theorem, 334

Boolean algebra of sets, 300

Borel's approximation theorem, 107, 205, 206, 243

Borel set, 297

Bounded

linear function, 83

operator, 323

set, 79, 257, 275

support, 293

- Box dimension, 207
 Braid, 35
 theory, 35
 Brownian motion, 125, 140
- C**
 Calabi conjectures, 258
 Cantor set, 15–17, 104, 105, 111, 112, 324, 325
 regular, 16–18
 Cardinality, 14, 149
 Catenary curve, 311
 Catenoid, 313
 Cauchy–Schwarz inequality, 132
 Cayley graph, 271
 Center density, 94
 Characteristic, 234
 Chebyshev constant, 118
 Chromatic number, 187, 201, 325
 Class BPP problem, 148
 Classification theorem, 136
 Class NP problem, 62–64, 174
 Class P problem, 61, 147
 Cliques
 number, 202
 problem, 61, 62
 Clock arithmetic, 28
 Closed
 set, 79, 275, 321
 subspace, 71
 surface, 258
 C^m norm, 307, 308
 C^1 norm, 155
 Combinatorial dimension, 210
 Commutative ring, 134
 Commutativity, 180, 194
 Compact
 set, 275
 surface, 257
 Complete
 network, 38
 set of integers, 192
 surface, 258
 Completely multiplicative function, 6
 Complex
 number, 9, 27, 34, 58, 112, 180, 240, 336, 343
 polynomial, 34
 rational function, 34
 Composite number, 144, 239
 Composition
 of functions, 19, 266, 267, 351
- of permutations, 394
 of rotations, 351
 Computable real number, 389
 Congruent, 255
 Conjugacy class, 395
 Conjugate
 elements, 394
 function, 55
 Connected
 graph, 141, 270
 set, 113
 $coNP$ -complete problem, 63
 Continued fraction, 407
 expansion, 115, 407
 Continuous time simple random walk, 274
 Contraction, 291, 315
 Convergent sum, 322
 Convex
 body, 79, 131
 function, 84, 262, 318, 320
 optimization, 321
 set, 79, 275, 318
 Coprime, 211
 Cosine, 276
 c_0 -space, 83
 Countable set, 396
 Critical
 line (RH), 344
 point, 31
 probability for percolation, 364
 site percolation, 365
 Cubic
 equation, 43, 249
 packing, 185
 ring, 136
 Curvature, 310, 311
 Gaussian, 312
 mean, 312
 principal, 312
 total, 313
 Cycle, 55, 141
 attracting, 56, 161
 Cyclotomic polynomial, 240, 241
- D**
 Decision problem, 61
 Decomposable form, 12
 Degenerate elliptic equations, 265
 De Giorgi's conjecture, 306
 Degree
 of a monomial, 10
 of a number field, 182, 196

- of a polynomial, 328
- of a rational function, 31
- of an algebraic number, 399
- Density**, 151, 165, 186, 347
- Derivative**, 81, 89, 155, 236, 303, 311, 318, 329, 338, 343
 - partial, 305, 308
 - second, 156
- Determinant**, 182, 402
- Diagonal Ramsey number**, 341
- Diameter**, 104
 - of a normed vector space, 171
- Dichotomy method**, 72
- Dictatorship rules**, 358
- Differentiable function**, 81
- Differential equation**, 303
- Diffusion**, 119
 - coefficient, 122
- Dimension**, 17, 206
 - box, 207
 - combinatorial, 210
 - Hausdorff, 17, 207
 - of a Banach space, 83
- Diophantine number**, 101, 105
- Dirichlet's approximation theorem**, 205, 206, 212, 243
 - two-dimensional, 246
- Dirichlet's theorem on primes in arithmetic progressions**, 333
- Discrete random variable**, 123
- Discriminant**, 182, 197, 402
- Distance**, 79, 176, 178, 283, 285
- Distortion**, 179
- Distributivity**, 194
- Divisor**, 239
- Domain**, 343
- d -regular graph**, 39
- Duality conjecture for entropy numbers of linear operators**, 133
- Duffin–Schaeffer conjecture**, 214
- Dynamical system**, 16, 53, 102, 161, 235
 - ergodic, 377
- Dyson's rank partition function**, 363

- E**
- Edge**, 141, 172, 187, 201, 253, 269, 325, 359
- Efficient algorithm**, 173
- Egyptian fraction**, 85
- Elliott–Halberstam conjecture**, 334
- Elliptic curve**, 43
- Embedded genus-one helicoid**, 314
- Embedding**, 179

- of groups, 396
- Empty set**, 377
- Enneper surface**, 159, 258
- Epigraph**, 318
- Equation**
 - cubic, 43, 249
 - degenerate elliptic, 265
 - differential, 303
 - Lebesgue–Nagell, 200
 - linear, 28, 42
 - Pell, 404
 - negative, 404, 405, 407
 - positive, 404, 405
 - Pell–Fermat, 404
 - quadratic, 28, 43
- Equidistributed trajectory**, 103
- Equivalence class**, 298
 - of rational functions, 33
- Equivalent rational functions**, 33
- Erdős' conjecture**, 288, 289
- Erdős–Graham conjecture**, 86, 88
- Erdős–Szemerédi problem**, 97
- Erdős–Turán conjecture**, 165, 166
- Ergodic**
 - direction, 103
 - dynamical system, 377
 - transformation, 377
- Erratic set**, 238
- Error correcting codes**, 93
- Euler's formula**, 114
- Even permutation**, 23, 352
- Eventually periodic**, 221
- Exceptional times for percolation**, 366
- Exhaustive submeasure**, 300
- Expander**, 39
 - family of, 39
- Expansion**, 270
 - coefficient, 270
- Expectation**, 373
- Expected value**, 3
- Exponential distribution**, 120

- F**
- f -almost every**, 214
- Family of expanders**, 39, 271
- Fatou set**, 115
- Fermat's Little Theorem**, 146
- Fibonacci numbers**, 198
- Field**, 26, 37, 45, 180, 194, 234
 - finite, 28, 45
 - of complex numbers, 27
 - of rational functions, 29, 45

- of rational numbers, 27
- of real numbers, 27
- quadratically closed, 29
- Fields medal, v
- Finite
 - field, 28, 45
 - group, 19, 271, 394
 - metric space, 178
 - simple groups, 26
 - type inequality, 12
- Finitely
 - additive measure, 299, 300
 - generated group, 59
 - presented group, 60, 62
- Fixed point, 54, 235
 - attracting, 54
- Forest, 143
- Form, 49
 - affine-linear, 384
 - irreducible, 52
- Fourier analysis, 175
- Fourier transform, 94
- Fractional part, 212, 245
- Fraenkel asymmetry, 296
- Fréchet differentiable function, 83
- Function
 - absolutely integrable, 378
 - admissible, 95
 - bounded linear, 83
 - completely multiplicative, 6
 - complex rational, 34
 - composition, 19
 - conjugate, 55
 - convex, 84, 262, 318, 320
 - differentiable, 81
 - Fréchet differentiable, 83
 - holomorphic, 343
 - identity, 266, 291
 - integrable, 378
 - inverse, 19, 266
 - invertible, 153
 - isoperimetric, 63
 - linear, 83, 266, 291, 314
 - Lipschitz continuous, 83, 264
 - lower-semicontinuous, 321
 - measurable, 378
 - odd, 314
 - of several variables, 305
 - periodic, 276
 - rational, 29
 - real rational, 31
 - regular, 56, 162
 - Schwartz, 128
- simple, 377
- stochastic, 57
- Fundamental
 - lemma of the Langlands program, v
 - parallelepiped, 349
 - parallelogram, 347
 - solution (Pell equation), 408
 - theorem of algebra, 336, 337
- G**
 - 2-to-2 games, 175
 - 2-to-2 Games Conjecture, 176
 - Γ -null sets, 84
 - Gaussian curvature, 312
- Generalized
 - arithmetic progression, 151
 - Riemann hypothesis, 332
- 2-generated group, 396
- Generating set
 - of rational points, 44
- Generic, 16
- Geometric
 - mean, 117
 - progression, 96
- $GL_2(\mathbb{Z})$ -equivalent, 136
- Golden ratio, 108, 205, 213, 244
- Good parameters, 152
- Graph, 39, 61, 172, 187, 201, 253, 269, 325, 359
 - colouring, 187, 201
 - proper, 187, 201
 - connected, 141, 270
 - d -regular, 39
 - k -colourable, 325
 - Paley, 341
 - perfect, 204
 - square of, 41
 - square-free, 327
 - theory, 141, 201, 253, 325
 - triangle-free, 326
- Group, 19, 23, 58, 133, 352, 394
 - 2-generated, 396
 - abelian, 23, 133
 - finite, 19, 271, 394
 - finite simple, 26
 - finitely generated, 59
 - finitely presented, 60, 62
 - infinite, 19
 - of bijections, 19, 23
 - of integers, 19, 23
 - of isometries, 59
 - of non-zero real numbers, 19

of two elements, 20
 permutation, 394
 simple, 24, 353
 symmetric, 23
 torsion-free, 397
 trivial, 394

H

Half-Turn Symmetric (HTS) matrix, 75
 Half-Turn-Symmetric ASM (HTSASM), 76
 Hall's conjecture, 6, 297
 Hardy–Littlewood maximal inequality, 92
 centered, 91
 Hasse–Minkowski theorem, 249
 Hasse principle, 249
 Hausdorff
 dimension, 17, 104, 112, 207
 f -measure, 214
 Heat conduction, 1
 Height of an algebraic number, 108
 Helicoid, 158, 314
 Hexagonal packing, 93, 184, 186
 Hilbert's problems, 344
 Hilbert–Waring theorem, 350
 Holomorphic function, 343
 Homeomorphic sets, 255, 325
 Homogeneous polynomial, 49
 Homothetic transformation, 315, 316
 Hopf's condition, 234
 Hyperbolic
 cosine, 311
 point, 237
 polynomial, 238, 329
 set, 238
 spiral, 313
 Hyperplane, 79, 306

I

Identity
 element, 23, 58, 194, 352, 394
 function, 266, 291
 operator, 322
 Imaginary number, 27
 Inclusion-exclusion principle, 287
 Induced subgraph, 204
 Inequality of finite type, 12
 Infinite
 group, 19
 order, 396
 Infinite-dimensional space, 70, 171
 Influence, 358

Inner product, 321
 Integers, 195, 398
 Integrable function, 378
 Integral, 88
 basis of a number field, 401
 group ring, 20
 ternary quadratic form, 136
 trace form, 402
 Integration, 303
 Interior, 79
 Interlacing polynomials, 330
 Interval exchange, 230
 Inverse, 59
 element, 352
 function, 19, 266
 matrix, 268
 vector, 71
 Invertible
 function, 153
 operator, 323
 transformation, 321
 Irrational number, 107, 212, 243, 389
 Irreducible
 action, 228
 form, 52
 fraction, 211
 polynomial, 239
 Isometry
 plane, 59
 Isomorphic
 fields, 196, 400
 groups, 20, 133, 396
 normed vector spaces, 71, 171
 rings, 21, 134
 vector spaces, 170
 Isomorphism, 170
 Isoperimetric
 deficit, 296
 function, 63

J

Jordan domain, 115
 Julia set, 112

K

k -colourable graph, 325
 Khinchin's theorem, 213, 244
 two-dimensional, 246
 Kissing number, 279, 281
 KMS theorem, 102–105
 Kolmogorov–Pollard entropy, 210
 Kronecker, L., 242

L

Lagrange's four squares theorem, 192, 232, 350, 351
 Lattice, 181, 346
 torus, 139
 triangular, 364
 Lebesgue differentiation theorem, 90, 91
 Lebesgue measure, 84, 104, 105, 113, 153, 213
 Lebesgue–Nagell equation, 200
 Leech lattice, 349
 Legendre transform, 320, 321
 Length of a vector, 168
 Level sets, 305
 Limit of a sequence, 71
 Line segment, 78
 Linear
 equation, 28, 42
 function, 83, 266, 291, 314
 operator, 323, 330
 transformation, 36, 321, 409
 Linearly independent
 polynomials, 337
 vectors, 335, 336
 Liouville's constant, 101, 244
 Lipschitz continuous function, 83, 264
 Little-o notation, 81, 83
 Littlewood's conjecture, 206, 207
 Locally connected set, 114
 Logarithmic capacity, 118
 Lower-semicontinuous function, 321
 L_p space, 285
 L_1 space, 378
 l^2 space, 71, 72, 322

M

Mahler's measure, 242
 Majority is stablest conjecture, 359, 360
 Matrix, 36, 67, 73, 267, 271, 293, 317, 402, 410
 alternating sign, 73
 half-turn symmetric, 75
 HTS, 75
 positive semi-definite, 68
 QTS, 75
 quarter-turn symmetric, 75
 square, 73
 symmetric, 67, 411
 transpose, 67
 unimodular, 268
 vertically symmetric, 74
 alternating-sign, 74

VS, 74

Matrix-positive polynomial, 68
 Max-cut problem, 359
 Mean curvature, 312
 Measurable
 function, 378
 set, 153, 377
 Measure, 300, 377
 finitely additive, 300
 space, 377
 Measure-preserving transformation, 153, 377
 Method of
 repeated squares, 146
 undetermined coefficients, 304
 Metric
 cotype, 286
 entropy, 210
 space, 178, 285
 finite, 178
 Millennium prize problems, 45, 62, 344
 Miller–Rabin primality test, 147, 148
 Minimal surface, 158, 258, 313
 Minkowski symmetrization, 77, 79
 Modular arithmetic, 268, 269
 Monic polynomial, 240
 Monomial, 10, 29, 49
 Monotone property, 327
 Multiplication by constant operator, 322
 Multiplicative
 Banach–Mazur distance, 171
 inverse, 194
 Multiplicity, 329

N

Natural numbers, 180, 398
 n -dimensional space, 70, 78, 131, 171, 297
 Negative numbers, 180
 Negative Pell equation, 404, 405, 407
 Newton's laws of motion, 375
 Nilmanifolds, 153
 Noise stability, 359
 Nonconventional averages, 153
 Non-ergodic direction, 103
 Non-nested arithmetic progressions, 362
 Non-trivial
 solution, 28
 subgroup, 24, 353
 zeros (of the zeta function), 344
 Norm, 71, 83, 153, 292, 322, 410
 C^1 , 155
 Normal

- plane, 312
- subgroup, 24, 353
- subset, 24
- Normed vector space, 71, 170
- NP -complete problem, 62
- NP -hard problem, 174
- Number field, 196, 399

- O**
- Obtuse triangle, 106
- Odd
 - antihole, 203
 - function, 314
 - hole, 203
- Open set, 113
- Operator, 322
 - bounded, 323
 - identity, 322
 - invertible, 323
 - linear, 323, 330
 - multiplication by constant, 322
 - shift, 322
- Order
 - of a group element, 395
 - of magnitude, 309
- Origin, 79
- Outer measure, 300

- P**
- Packing
 - cubic, 185
 - hexagonal, 93, 184
 - sphere, 94, 185
 - square, 183
- p -adic field, 249
- p -admissible weight, 265
- Painlevé–Hadamard Principle, 58
- Pairwise comparison paradox, 357
- Paley graph, 341
- Palis conjecture, 16, 17
- Parameter, 367, 383
- Partial derivative, 305, 308
- Partitions, 361
- Path, 364
- $P = BPP$ conjecture, 148
- PCP theorem, 175
- Pell equation, 404
 - negative, 404, 405, 407
 - positive, 404, 405
- Pell–Fermat equation, 404
- Percolation, 364
- Perfect graph, 204
- Period, 407
- Periodic
 - expansion, 407
 - function, 276
 - point, 237
 - attracting, 237
- Permutation, 23, 351
 - even, 23, 352
 - group, 394
- Pigeonhole principle, 245
- Plane
 - curve, 313
 - isometry, 59
 - reflection, 59
 - rotation, 36, 59, 351
- (1, p)-Poincaré inequality, 265
- Poincaré conjecture, vi
- Poincaré inequality, 265
- Poisson distribution, 273
- Poisson process, 98
- Polar body, 132
- Pólya’s Theorem, 124
- Polynomial, 11, 29, 239, 328, 336
 - algorithm, 61, 173
 - complex, 34
 - cyclotomic, 240, 241
 - family, 369
 - homogeneous, 49
 - hyperbolic, 238, 329
 - irreducible, 239
 - matrix-positive, 68
 - monic, 240
 - quadratic, 161, 304
 - reducible, 239
 - stable, 329, 331
 - symmetric, 69
- Positive Pell equation, 404, 405
- Positive semi-definite matrix, 68
- Powers (in a group), 353
- Prime number, 144, 224, 239, 259, 332, 342, 379
 - theorem, 167, 196, 224, 261, 332, 334, 382
- PRIMES, 147
- Primitive solution, 370
- Principal curvatures, 312
- Probability, 3
 - of large deviations, 4
 - of moderate deviations, 4
 - theory, 2
- Product
 - of matrices, 36, 67, 267

- zig-zag, 39
- Projection, 291
- Proper graph colouring, 187, 201
- Properly embedded
 - curve, 313
 - surface, 159, 313
- P versus NP , 62
- Pythagoras' Theorem, 21, 106, 211, 367, 389

- Q**
- Quadratic
 - equation, 28, 43
 - non-residue, 7
 - polynomial, 161, 304
 - residue, 7
 - ring, 135
- Quadratically closed field, 29
- Quantitative noise sensitivity estimates, 366
- Quarter-Turn Symmetric ASM (QTSASM), 76
- Quarter-Turn Symmetric (QTS) matrix, 75
- Quartic ring, 136

- R**
- Ramsey theorems, 73
- Random
 - band matrix, 411
 - matrix, 411
 - variable, 293, 411
 - discrete, 123
 - subgaussian, 293
 - walk, 137, 274
- Rank
 - of a partition, 362
 - of an elliptic curve, 45
- Rate constant, 126
- Rational
 - function, 29
 - complex, 34
 - number, 107, 180, 194, 205, 212, 221, 298, 398
 - point, 43, 53
 - of infinite order, 44
- Real
 - algebraic geometry, 338
 - enumerative geometry, 34
 - number, 180, 194
 - rational function, 31
- Reducible
 - action, 228
 - polynomial, 239
- Reflection
 - through a hyperplane, 79
 - through a line, 59
 - through a point, 291
- Regular
 - Cantor set, 16–18
 - function, 56, 162
- Relatively prime, 247, 333, 381
- Renormalization, 58
- Riemann hypothesis, 344, 345
- Riemann zeta function, 344
- Right recursively enumerable, 390
- Ring, 20
 - commutative, 134
 - cubic, 136
 - integral group, 20
 - ring
 - of rank n , 134
 - quadratic, 135
 - quartic, 136
 - trivial, 136
- Root, 240, 336
- Rotation
 - of a plane, 36, 59, 291, 351
 - of space, 36
 - shuffle, 228
- Roth's theorem, 166, 251
- Ruled surface, 160
- r -uniform hypergraphs, 253

- S**
- Scalar product, 306
- Schinzel–Zassenhaus conjecture, 242
- Schwartz function, 128
- Second derivative, 156, 304, 311, 338
- Separable Banach space, 172
- Set of generators
 - of a group, 59
 - of Cayley graphs, 271
 - of matrices, 269
- Set of measure 0, 58, 84, 90, 104, 105, 113, 163, 213, 376
- Shift
 - of finite type (SFT), 391, 392
 - operator, 322
- Shift of Finite Type (SFT), 391, 392
- Siegel disk, 115
- Sieve of Eratosthenes, 144
- Simple
 - closed curve, 158
 - curve, 313
 - function, 377
 - group, 24, 353

- solution, 50
 - Simply connected surface, 160
 - Simultaneous differentiability, 84
 - Six-vertex model, 74
 - Smale's problems, 238
 - Sophie Germain conjecture, 386–388
 - SoS method, 66
 - Space
 - of continuous functions, 171
 - of convergent sequences, 171
 - rotation, 36
 - Spectrum, 9, 323
 - Sphere packing, 94, 185
 - Square
 - element of a group, 353
 - of a graph, 41
 - packing, 183
 - Square-free
 - graph, 327
 - integer, 370, 404
 - Stable
 - intersection, 17
 - polynomial, 329, 331
 - Standard double bubble, 49
 - Stevenhagen's conjecture, 408
 - Stochastic
 - function, 57
 - set, 163
 - Stretching, 315
 - Subgaussian random variable, 293
 - Subgraph
 - induced, 204
 - Subgroup, 23, 64, 352, 396
 - non-trivial, 24, 353
 - normal, 24, 353
 - trivial, 24, 353
 - Submeasure, 300
 - exhaustive, 300
 - Subset sum problem, 61, 62
 - Subspace, 71
 - closed, 71
 - Subtree, 142
 - Sum
 - of digits, 381
 - of matrices, 67
 - of squares method (SoS), 66
 - Sum-product conjecture, 97
 - Sums-of-squares formula, 235
 - Superpolynomial improvement, 342
 - Surface of finite topology, 258, 314
 - Symmedian point, 233
 - Symmetric
 - body, 132
 - difference, 295
 - group, 23
 - matrix, 67, 411
 - polynomial, 69
 - transformation, 321
 - Syndetic set, 150
 - Szemerédi's regularity lemma, 253
 - Szemerédi's Theorem, 151, 252
 - multidimensional, 252, 253
- T**
- Ten martini problem, 325
 - Topological entropy, 391, 392
 - Topology, 160, 255
 - Torsion-free group, 397
 - Total curvature, 313
 - Totally Asymmetric Simple Exclusion Process (TASEP), 371
 - Trace, 400
 - Transcendental number, 109, 222
 - Transfinite diameter, 118
 - Transformation
 - affine-linear, 383
 - ergodic, 377
 - homothetic, 315, 316
 - linear, 409
 - measure preserving, 377
 - Translate, 149, 392
 - Transpose matrix, 67
 - Tree, 141
 - Triangle, 61
 - inequality, 71, 83, 170, 176, 285
 - rule, 168
 - Triangle-free graph, 326
 - Triangular lattice, 364
 - Trivial
 - group, 394
 - ring, 136
 - solution, 28
 - subgroup, 24, 353
 - zeros (of the zeta function), 344
 - Twin primes, 224
 - conjecture, 225, 332, 333, 335, 386–388
 - Two-dimensional asymmetric simple exclusion process, 121
- U**
- u -invariant, 29
 - Uncountable set, 396
 - Undecidable problem, 65
 - Uniformly embedded, 285
 - Uniform Spanning Forest (USF), 143

Unimodal map, 164
 Unimodular matrix, 268
 Upper density, 151, 165, 186

V

Variance, 124, 293, 373
 Vector, 67, 71, 168, 335, 346
 space, 71, 170
 normed, 71, 170
 of continuous functions, 72
 Vertex, 141, 172, 187, 201, 253, 269, 325,
 359
 cover, 172
 Vertically symmetric
 alternating-sign matrix, 74
 matrix, 74
 Vertically Symmetric Alternating-Sign
 Matrix (VSASM), 74
 Vitali set, 298, 299
 VS matrix, 74

W

Weakly mixing, 230
 Weighted
 average, 264
 length, 264
 Whitney's extension theorem, 157
 Wiener
 process, 4, 125
 sausage, 4, 126
 Witness, 147
 Word, 59, 353
 problem, 60, 62–64
 Wronskian, 338

Z

Zero vector, 71
 Zig-zag product, 39